DOCUMENT RESUME

ED 262 066 TM 850 541

AUTHOR	Fairbank, Benjamin A., Jr.
TITLE	Equipercentile Test Equating: The Effects of Presmoothing and Postsmoothing on the Magnitude of Sample-Dependent Errors.
INSTITUTION	Performance Metrics, Inc., San Antonio, TX.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Tex. Manpower and Personnel Div.
REPORT NO	AFHRL-TR-84-64
PUB DATE	Apr 85
NOTE	168p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Adults; *Equated Scores; Equations (Mathematics); *Error of Measurement; Psychometrics; *Regression (Statistics); *Sampling; Simulation; Statistical Analysis
IDENTIFIERS	*Armed Services Vocational Aptitude Battery; *Equipercentile Equating; Hypergeometric Distribution; Jackknifing Technique

ABSTRACT

The effectiveness of 19 methods of smoothing was investigated as those methods apply to the equipercentile method of test equating. Seven methods involved smoothing the score distribution before the tests were equated (presmoothing). Seven involved smoothing the resultant points after the equating (postsmoothing). Five methods involved combining presmoothers and postsmoothers. The results of smoothing were evaluated by comparing smoothed and unsmoothed equatings with large sample equating or other criterion equatings. The data that were used include the results of test simulations and results of administrations of military selection tests. Measure of average absolute deviation, average signed deviation, and root mean square deviation were calculated. Jackknifing was also used to estimate standard errors of equatings. The nonlinear presmoothers were less effective than a presmoother based on negative hypergeometric distribution. Among the postsmoothers, the most promising was a technique using cubic smoothing splines. Combining presmoothers and postsmoothers was not notably more effective than either smoother alone. It is suggested that the family of smoothing functions based on the negative hypergeometric be more fully investigated. (Author)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

AFHRL-TR-84-64

AIR FORCE

HUMAN RESOURCES

ED262066

EQUIPERCENTILE TEST EQUATING: THE EFFECTS OF PRESMOOTHING AND POSTSMOOTHING ON THE MAGNITUDE OF SAMPLE-DEPENDENT ERRORS

By

Benjamin A. Fairbank, Jr.

Performance Metrics, Inc.
5825 Callaghan Road, Suite 225
San Antonio, Texas 78228

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

April 1985
Final Report for Period August 1983 - March 1984

Approved for public release; distribution unlimited.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
AFHRL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LABORATORY

TM 850 541

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

BEST COPY AVAILABLE

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

TONI WEGNER, Captain, USAF
Contract Monitor

NANCY GUINN, Technical Director
Manpower and Personnel Division

ANTHONY F. BRONZO, Jr., Colonel, USAF
Commander

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFHRL-TR-84-64

6a. NAME OF PERFORMING ORGANIZATION
Performance Metrics, Inc.

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Manpower and Personnel Division

6c. ADDRESS *(City, State and ZIP Code)*
5825 Callaghan, Suite 225
San Antonio, Texas 78228

7b. ADDRESS *(City, State and ZIP Code)*
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
Air Force Human Resources Laboratory

8b. OFFICE SYMBOL *(If applicable)*
HQ AFHRL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F41689-83-C-0042

8c. ADDRESS *(City, State and ZIP Code)*
Brooks Air Force Base, Texas 78235-5601

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
62703F	7719	18	38

11. TITLE *(Include Security Classification)* Equipercentile Test Equating: The Effects of Presmoothing and Postsmoothing on the Magnitude of Sample-Dependent Errors

12. PERSONAL AUTHOR(S)
Fairbank, Benjamin A. Jr.

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM Aug 83 TO Mar 8[illegible]

14. DATE OF REPORT *(Yr., Mo., Day)*
April 1985

15. PAGE COUNT
166

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.
05	09	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
4253H Twice
Armed Services Vocational Aptitude Battery (ASVAB)
average absolute deviation
average signed deviation
cubic regression

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The effectiveness of 19 methods of smoothing was investigated as those methods apply to the equipercentile method of test equating. Seven methods involved smoothing the score distribution before the tests were equated (presmoothing). Seven involved smoothing the resultant points after the equating (postsmoothing). Five methods involved combining presmoothers and postsmoothers. The results of smoothing were evaluated by comparing smoothed and unsmoothed equatings with large sample equating or other criterion equatings. The data that were used include the results of test simulations and results of administrations of military selection tests. Measure of average absolute deviation, average signed deviation, and root mean square deviation were calculated. Jackknifing was also used to estimate standard errors of equatings. The nonlinear presmoothers were less effective than a presmoother based on negative hypergeometric distribution. Among the postsmoothers, the most promising was a technique using cubic smoothing splines. Combining presmoothers and postsmoothers was not notably more effective than either smoother alone. It is suggested that the family of smoothing functions based on the negative hypergeometric by more fully investigated.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Nancy A. Perrigo
Chief, STINFO Office

22b. TELEPHONE NUMBER *(Include Area Code)*
(512) 536-3877

22c. OFFICE SYMBOL
AFHRL/TSR

DD FORM 1473, 83 APR
EDITION OF 1 JAN 73 IS OBSOLETE.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

Item 18 (Continued)

cubic smoothing splines
equipercentile equating
five-point moving medians
five-point moving weighted averages
five-point moving weighted averages with root transformation
jackknifing
linear regression
logistic ogive
military enlistment test
negative hypergeometric
orthogonal regression
postsmoothing
presmoothing
psychometrics
quadratic regression
root mean square deviation
simulations
smoothing
standard errors
test equating
three-point moving medians
three-point moving weighted averages

Summary

The Air Force requires effective methods for test equating. Among the tests which must be equated are the various forms of the Armed Services Vocational Aptitude Battery (ASVAB). Equipercentile test equating is typically used to equate the different forms of the ASVAB to earlier forms and to each other. Increases in the accuracy of equipercentile test equating may be achieved by increasing the size of the samples of examinees. The purpose of the present effort was to determine whether statistical smoothing could also increase the accuracy of equating.

Two classes of simple smoothing methods are of interest - presmoothing of the score distributions and postsmoothing of the equipercentile points. A third class of smoothing methods, called combined smoothers, involved both presmoothing and postsmoothing. The research used three methods to investigate fourteen simple smoothers and five combined smoothers. The first method used simulations based on a theory of ability testing. Simulated tests were developed to mimic statistical aspects of ASVAB subtests. Those tests were equated with and without smoothing and the results were evaluated. The second and third methods used existing operationally obtained data. In the second method, very large samples of examinees were used to establish highly accurate equatings, then smaller samples were drawn and equated with and without smoothing. The third method of investigation used the statistical jackknife, a general purpose statistical tool, to estimate standard errors.

Negative hypergeometric presmoothing was clearly more effective than the other presmoothers. Two of the postsmoothers were somewhat more effective than the other postsmoothers. The negative hypergeometric presmoother resulted in a reduction of approximately ten percent in one measure of equating error; its use would correspond in effectiveness to an increase of approximately twenty percent in the size of the samples used for equating. The effective postsmoothers were (1) orthogonal regression, which was more effective than ordinary least squares linear regression, and (2) the use of cubic smoothing splines, which was the most effective of the postsmoothers. No postsmoother was as effective as presmoothing with the negative hypergeometric. Combining presmoothers and postsmoothers did not result in an improvement beyond that obtained with the more effective of the combined pair used alone.

Modest but significant gains in the accuracy of equipercentile test equating may be achieved through the use of negative hypergeometric presmoothing.

Preface

This research and development (R&D) effort was conducted under Project 7719, Force Acquisition and Distribution System; Task 771918, Personnel Qualification Tests. This is part of a continuing program of R&D on methods which may be used to improve the assessment of personnel qualifications. The specific work unit, 77191838, is titled Equating Smoothing.

The author wishes to express his appreciation to Dr. Malcolm Ree and Capt. Toni Wegner of the Manpower and Personnel Division of the Air Force Human Resources Laboratory. Their discussions, suggestions, and support were essential to this effort. In addition, suggestions and insights provided by Dr. Michael Levine and Dr. Mark Reckase in the early stages of the research were most valuable and are appreciated.

TABLE OF CONTENTS

		Page
I.	Introduction	1
II.	Methods	8
III.	Results	20
IV.	Discussion	46
	References	52
	Appendix A: Technical description of simulated tests	55
	Appendix B: Figures showing the effects of the smoothing procedures	65

LIST OF FIGURES

Figures		Page
1	Test information curves (upper panel) and test characteristic curves (lower panel) for the simulated tests of length 15.	12
2	Test information curves (upper panel) and test characteristic curves (lower panel) for the simulated tests of length 30.	13
3	Test information curves (upper panel) and test characteristic curves (lower panel) for the simulated tests of length 50.	14
B-1 through B-95		66-160
	Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equatings.	

LIST OF TABLES

Tables		Page
1	Mean Measures of Deviations for Unsmoothed Equatings	22
2-20	Summaries of the Averaged Effects of Smoothers	23-41
21	Standard Errors of Unsmoothed Equating Estimated by Three Methods	42
22	Proportional Change in Standard Errors as a Result of 19 Smoothing Methods	43
A-1 through A-7		56
	Technical Descriptions of Simulated Tests	

Equipercentile Test Equating: The Effects of Presmoothing and Postsmoothing on the Magnitude of Sample-Dependent Errors

I. INTRODUCTION

Test Equating

Test equating is the process of finding which scores on two or more similar tests correspond to the same level of ability in the population of examinees. In principle, when two tests have been equated, either can be used with equal confidence to measure ability. Test results will then be on the same scale and any examinee's expected score will not be affected by the form of the test administered. The tests under consideration in this report are four-option, multiple-choice tests that are scored on the basis of the number of correct responses. This report addresses certain methodological issues which arise in the process of equating. Before delineating those issues, however, the process of equating will be put into the larger context of testing in general.

The need for test equating arises as a result of many considerations. It is often valuable to have more than one version or form of a test. When more than one version or form of a test is available, the particular form taken by an examinee should not affect the examinee's expected score. In other words, there should be no advantage or disadvantage associated with taking one form of a test rather than another form of the same test.

The need for more than one form of a test may arise from any of a number of considerations, including testing policies which allow the examined individual to be reexamined with a different form of the test. Alternatively, the need may be a consequence of security considerations. If there are several forms of a test in existence, then a compromise of the security of one test form does not compromise the entire testing program.

The replacment of operational tests requires equating when the scores on the new tests are to be used in the same predictive or evaluative equations or in the same manner as were the old scores. The need for replacing operational tests can be due to changes in the characteristics of the tested population, in the effectiveness of some of the test items, or in the needs of the testing agency. Test replacement policies can also be a response to the possibility of test compromise, or a breach of test security. As a test is operational for longer and longer periods, the chances increase that the test may no longer be secure against unauthorized disclosure and, hence, that scores may no longer represent ability.

The Armed Services conduct the largest testing program in the nation. The Armed Services Vocational Aptitude Battery (ASVAB) is administered annually to all applicants (about one million) for enlistment in the Armed Services, as well as to about one million high school and post-secondary school students. The applicants' acceptance into or rejection by the Services is governed in part by the results of the tests, as is their assignment to particular specialities once they are in the Service. The success and security of the testing program are thus important for the continued effective use of personnel in the Services. All of the considerations mentioned above which lead to the need for test equating are present in the ASVAB testing program. Both the requirement for several forms and the requirement for periodic replacement of forms lead to the necessity of test equating. It is therefore important that effective methods of test equating be available to the psychometric community within the Armed Services.

Test equating may be carried out in any of a large number of different ways, some of which are of recent origin and are technically sophisticated, and some of which have been in use for several decades (see Holland & Rubin, 1982). This report addresses only equipercentile test equating as applied to two equivalent groups (Angoff, 1971). A brief description of equipercentile test equating is given here; a more complete description may be found in Angoff (1971). The logic of equipercentile equating is based on the concept that the ability of examinees who take tests may be used to calibrate or equate the tests. If two groups of examinees have identical distributions of abilities, and if one group takes one test and the other group takes another test, then corresponding percentile points in the two groups will correspond to equal abilities, and those points can be used to establish corresponding, or equated, scores on the two tests. For example, if individuals in the 20th percentile received a score of 23 on one test, and if the individuals in the 20th percentile of the other group received a score of 25 on the other version of the same test, then a score of 23 on the first test is said to be equated to a score of 25 on the second test. In practice, equated scores are not usually given for every percentile point, but rather for every obtainable test score on each of the tests. One can then convert from either test to the other with equal facility. The terms "reference test" and "experimental test" are used to indicate, respectively, the test whose score metric is to be used for the results of both tests, and the test whose score is to be converted to the units of the other test. For example, if an existing test known as Form K is to be replaced by a similar test known as Form M, Form K would be the reference test and Form M would be the experimental test.

In order for equating to be accurate (i.e., for the tests to be used interchangeably with no advantage or disadvantage associated with the taking of either test) two conditions must be met. First, the two groups of examinees used for equating must be equivalent, and second, the two tests must measure the same trait equally reliably. The equivalence of the two groups is usually met in practice by having one group divided at random into two smaller groups. The question of whether two tests are sufficiently similar is more difficult. Lord (1980) demonstrates that two tests cannot be equated unless they are either perfectly reliable (an impossibility), or are strictly parallel, in which case they would not need to be equated. In practice, however, it is possible to equate highly similar tests, sometimes called "roughly parallel" tests, by the equipercentile method in such a way that the errors of equating are very small in comparison with other errors associated with testing (e.g., the errors of measurement arising as a consequence of the unreliability of tests, and particularly the inherent lower reliability of short tests). In any case, although there may be some purposes to which it would be misleading to put equated scores, Lord (1980) points out that if scores are equated by the equipercentile method, then when equated cutting scores are used, the different equated forms will result in the selection of the same proportion of examinees on all forms of the test, except for errors related to sampling in the equating process or to the particular examinees tested operationally.

As with any procedures having the goal of estimating population characteristics based on data obtained from a sample, there are always sample-dependent errors present in test equating. If an equipercentile equating were to be done twice with similar samples, the results would differ. The extent of such differences has been estimated by Lord (1982) and their magnitudes appear as the standard errors of equipercentile equating. As with all standard statistical procedures, the size of the expected errors decreases linearly with the square root of the sample size. It is thus operationally impractical to reduce errors beyond a certain amount by increasing sample sizes. For example, decreasing the error to one-fourth the size of the error associated with a given sample size would require using a sample 16 times the size of the original sample. As a consequence, practitioners of equipercentile test equating have looked for other ways to reduce equating errors. They have most frequently used the methods of smoothing.

Smoothing

Two general classes of smoothing methods are defined here. A third class is made up by combining a smoothing method from the first class with one from the second class. First, presmoothing is defined as the process of smoothing the observed score frequency distributions prior to the equating. Second, postsmoothing is defined as the process of smoothing the equipercentile points after equating. Third, combined smoothings involve presmoothing and postsmoothing applied consecutively. The common intent of all three smoothing methods is to remove small sample-dependent fluctuations from the nonsmoothed equatings so that the small sample equatings will more nearly approximate the asymptotic equatings, or those which would result from the use of samples so large that the sample-dependent errors approach zero. The extent to which the various methods achieve this common intent is investigated by this research. Seven presmoothing methods, seven postsmoothing methods, and five combined smoothing methods were used as follows:

A. Presmoothing Methods
   1. 3-point moving medians
   2. 5-point moving medians
   3. 3-point moving weighted averages
   4. 5-point moving weighted averages
   5. 5-point moving weighted averages with root transformation
   6. 4253H Twice
   7. negative hypergeometric

B. Postsmoothing Methods
   1. linear regression
   2. quadratic regression
   3. cubic regression
   4. orthogonal regression
   5. logistic ogive
   6. cubic splines
   7. 5-point moving weighted averages

C. Combined Smoothers
   1. negative hypergeometric + orthogonal regression
   2. negative hypergeometric + quadratic regression
   3. negative hypergeometric + 5-point moving weighted averages
   4. 3-point moving weighted averages + 5-point moving weighted averages
   5. negative hypergeometric + cubic splines

Presmoothing

Presmoothing methods are based on the concept that an observed data point in a sequence of points shows the combined effect of an underlying systematic relation among the points and sample-specific fluctuation or error of observation. If each point were replaced by a value jointly determined by the point replaced and the vicinal points, then the influence of the error of observation should be reduced, and the influence of the underlying regular function should be enhanced.

Six of the seven presmoothing methods used in this study are general-purpose methods which were developed for the smoothing of sequences of observations such as time series data (Keats & Lord, 1962; Tukey, 1977; Velleman, 1980; Velleman & Hoaglin, 1981). Detailed technical descriptions of the methods are available in the references cited; short descriptions are provided here.

Moving medians and moving averages were used for presmoothing, as were a combined or compound presmoother and a presmoothing method based on a particular model of test scores.

Of the seven methods of presmoothing the score distributions, three are described by Tukey (1977). In the first method, frequency distributions are smoothed by moving medians of span three. Smoothing by moving medians of span three involves replacing each observed frequency with the median of three frequencies: that of the score of interest, the frequency associated with the next lower score, and that associated with the next higher score. The end values of the distribution, those corresponding to scores of 0 and perfect scores, are not smoothed because they have only one neighboring value, and thus cannot be smoothed effectively by moving medians. Moving medians of span five are found analogously, except that each frequency is replaced with a value which is the median of the frequency of interest, the two preceding frequencies, and the two following frequencies. The end points are not smoothed, but the next-to-end points are, by convention, replaced by the smoothed values found by smoothing by medians of span three.

Presmoothing by three-point moving weighted averages is analogous to three-point moving medians, but instead of replacing each point in the raw frequency distribution with its median, it is replaced with a value that is calculated by taking the sum of twice the point being smoothed, the previous point, and the following point, then dividing the result by four. This is equivalent to using weights of 1, 2, and 1. The weights 1, 2, and 1 are chosen to give the point being smoothed a weight equal to the surrounding points in determining the smoothed value. Clearly, any other weighting is possible, from one in which weights of 0, 1, 0 correspond to no smoothing, to one in which weights of 1, 0, 1 correspond to a smoothing in which a point's surrounding points completely determine its value. Again, the end values are not smoothed. Five-point moving weighted averages are found by taking the raw frequencies five at a time and replacing each frequency with a weighted average of the frequency and the four surrounding values. The weighting function is one recommended by Angoff (1971), and weights the five points by the factors -3, 12, 17, 12, -3, and divides the resulting sum by 35. The recommended weights allow linear, quadratic, and cubic components of the curve to be unaffected by the smoothing process. The end frequencies are not smoothed, but the next-to-end frequencies are smoothed by the three-point moving weighted average using weights of 1, 2, 1. The five-point moving weighted average with root transformation is identical to the five-point moving weighted average, except that before the smoothing is applied, all of the frequency values are transformed by taking their square root. The square roots are then smoothed. Following the smoothing, the inverse transformation, a squaring, is applied. The use of the square root transformation has the effect of decreasing the influence of larger values relative to the effect of the same smoother without the square root transformation. As a result, if a frequency is higher than surrounding frequencies, it is more effectively reduced with the root transformation. Conversly, if a frequency is lower than surrounding frequencies, it is more effectively raised to the surrounding values when the root transformation is not used. At the range of frequencies reported here, however, the differences are very slight.

The sixth smoother is a combination of smoothers proposed by Velleman (1980). Designated as 4253H Twice, it is a complex method which requires the successive application of four different smoothers, including moving medians of spans four, five, and three, then the finding of the differences between the smoothed and unsmoothed distributions, the smoothing of that sequence of differences by the same compound method, and, finally, adding the smoothed differences back into the smoothed distribution. The smoothing by medians of span 4 results in smoothed values which correspond to points between the originally smoothed points. Thus, smoothing points n, n+1, n+2, and n+3 results in a value corresponding to a point between points n+1 and n+2. The step designated by "2" in 4253H is required to bring each smoothed value back to its proper association with the point being smoothed. Details are given in Tukey (1977) and Velleman and Hoaglin (1981).

The final presmoothing method (see Keats & Lord, 1962; also Lord & Novick, 1968, pp. 515-520) is one devised explicitly for smoothing or fitting frequency distributions of test scores. The distribution is the negative hypergeometric, whose appropriateness is derived from a binomial error model of test scores. The model assumes several technical conditions, one of which is equivalent to the assumption that all of the items on the test whose score distribution is being fit are equally difficult. That condition is known to be false in the case of the ASVAB, as well as for most other tests, but the fit of the negative hypergeometric is still good enough to make it promising for further study (Keats & Lord, 1962).

Postsmoothing

Equipercentile equating, as described earlier in this section, starts with tables which show the frequency of each score in the samples tested for each of two tests and ends in a table which associates with each score on one test a score on the other test. An integer score on one test is usually found to correspond to a non-integer score on the other test; the non-integer score may be estimated by linear interpolation. A plot of the score pairs shows a monotonically nondecreasing function whose form depends on characteristics of the sample and characteristics of the two tests being equated.

Postsmoothing is the process of passing a straight line or a curve among the points which define the equipercentile relationship. The equated scores are then determined by the resulting function. Postsmoothing methods have traditionally required the practitioner to judge where to pass a curve through a set of points (Angoff, 1971). In place of the use of a draftsman's French curve or analogous drawing aid, a number of analytic postsmoothing methods have been developed. Seven such methods were investigated here. They were chosen on the basis of a number of considerations including practicality of implementation, frequency of use in the past, and the extent to which the methods appeared promising based on the literature.

The simplest equation which may be fit to the points resulting from an equipercentile equating is a straight line. This study investigated two different straight lines: that defined by conventional least squares and that defined by orthogonal regression. The conventional least squares procedure minimizes the sum of the squared vertical deviations from the line. In effect, the scores on the experimental test are considered to be known without error, and the line which best fits the equipercentile equivalents on the reference test is found. In orthogonal regression (Madansky, 1959), the quantity minimized is not the sum of the squared deviations parallel to the y-axis, but, rather, the sum of the squared deviations when those deviations are taken in a direction perpendicular to the regression line. Orthogonal regression is appropriate when the variables represented on both axes are subject to measurement error, and neither can properly be considered the dependent or independent variable. This is frequently the case in test equating, for two reasons. First, such an equating can be used to convert scores from either test to the other. It is thus dissimilar to a least squares regression equation in which the regression of y on x is rarely the same as that of x on y. Second, there are usually similar amounts of error associated with the reference and the experimental test. The first two postsmoothing methods, then, are straight lines fit by conventional regression and by orthogonal regression. When conventional regression is used, the independent variable is the set of scores on the experimental test ranging from the lowest observed score to the highest observed score. The dependent variable is made up of the equipercentile points.

Only under certain circumstances is it possible to fit resulting points well with a straight line. A straight line is appropriate if the two tests have the same skewness and kurtosis. The positioning and slope of the straight line will compensate for differences in means and standard deviations in the two tests. If there is a curvilinear component to the relationship defined by the equipercentile equating, then it must be fit by a curvilinear function. Quadratic and cubic functions are commonly used to fit such curves. This investigation considered quadratic and cubic best-fitting (criterion of minimum least squares deviations) smoothing curves. Quadratic curves can fit points whose best-fitting line is concave either upward or downward, whereas cubic equations can fit curves with an inflection point, so that part of the curve is concave upward and part of it is concave downward. The use of quadratic or cubic postsmoothing functions can result in nonmonotonic functions in which there is a part of the smoothing function at which an increase in the score on the experimental test results in a decrease in the equated score. Such reversals are artifacts of the fitting process and when they occur are corrected to monotonicity. The correction is made by forcing each score to be greater than or equal to the preceding score. Such correction is rarely needed. The third and fourth postsmoothing methods, then, were quadratic and cubic regression functions, fit by the method of least squares as modified by the requirement of monotonicity.

In some equatings it is observed that the equipercentile equating function is relatively flat at both of its ends and steeper in the middle. Such a shape can be fit by

a cubic curve, but it can also be fit by a logistic ogive, a curve defined by the equation

$$Y = A + \frac{B - A}{1 + \exp(-C(X - D))}$$

where A, B, C, and D are fitted constants. The points resulting from equipercentile equating were fit by a logistic ogive, the fifth postsmoothing method.

All of the smoothing methods mentioned above have associated with them the disadvantage that they impose a function of a given form on the data, even if it is not appropriate. Such a procrustean requirement is contrary to the rationale of smoothing, especially when the function is not appropriate in shape to the points to which it is to be fit. The sixth and seventh postsmoothing methods do not define the shape of the function in advance of the fitting.

The sixth function fit to the points was not a continuous function, but rather a smoothing of the discrete resulting points. The smoothing function replaces each point with a point which is the weighted average of the point being replaced and the four surrounding points. The method is that of five-point moving weighted averages, as described earlier. The equating requires interpolation between the resultant points.

The final postsmoothing function was used by Kolen (1983), who obtained good results by fitting cubic smoothing spline functions to the points resulting from the equipercentile equating. A smoothing spline differs from an interpolating spline in that the latter is constrained to pass through exactly known points, while the former is conceived of as passing among approximately known points. As used by Kolen, a cubic smoothing spline for N points (in the present case, an equipercentile equating of two N-item tests) is a set of N-1 cubic functions, each of which takes as its domain the interval from the I-th point to the (I + 1)th point on the X-axis. The range and specific form of the function are determined by the data in the interval. The cubic functions come together with the same function value and slope (or derivative) at each of the interior N-2 points, which are called ducks or knots in the language of spline fitting. The former term, ducks, is used in this report. The resulting curve can be of almost any differentiable shape, because the individual cubic fittings are independent of each other and can follow the shape of the function defined by the points to be smoothed.

Combined Smoothers

The use of a presmoothing technique does not preclude the use of a postsmoother. In order to determine whether or not presmoothing and postsmoothing employed consecutively would have benefits beyond those due to either method alone, five combinations of presmoothing and postsmoothing were investigated. Presmoothing with the negative hypergeometric was combined with four different postsmoothers, orthogonal regression, five-point moving weighted averages, cubic splines, and quadratic smoothing. Finally, presmoothing by the method of three-point moving weighted averages was combined with postsmoothing by means of five-point moving weighted averages.

Objectives

The aim of the present effort was to evaluate the effects of various different methods of presmoothing, postsmoothing, and combined smoothings on the accuracy of test equating. The study was exploratory in nature, designed to determine which methods hold the most promise for operational use. Detailed confirmatory scrutiny of the efficacy of promising methods will have to await future attention. The following section describes the methods used to determine the accuracy of equating and to apply the methods to simulated and to operational tests.

II. METHODS

General Plan

The plan underlying this investigation was to use three different approaches to determine the effectiveness of each of 14 unitary smoothing methods and five combined smoothing methods. The first approach used simulated tests and examinees; the second and third used data from tests administered to examinees under operational conditions. The advantage of simulated tests and examinees is that all quantitative aspects of the tests and examinees are completely specified, and it is possible to know in advance the results of theoretically errorless equatings or those equatings which are unaffected by sample-dependent errors. Operational data, of course, have the advantage that they are obtained under conditions typical of the ones under which smoothing methods would be used. The data are not based on an ideal model, as are the data from simulations; rather, they contain all of the departures from theory that may be found in operational test settings.

The first of the three methods of evaluation involved comparing each of the smoothed equatings with a known errorless equating. The known errorless equating was based on a method that yielded results typical of an equating using an infinitely large sample. The method requires deriving a distribution of expected score frequencies, the distribution being that which would result from administering the test to a sample so large that the observed proportions at each score were observed essentially without error. The results of the simulated test administrations were used for that method. The second method was a similar comparison of sample and criterion equatings, but in place of data based on simulations and on an errorless equating, the comparison used operationally obtained data and an equating based on an unusually large sample size. The third method was to use the statistical jackknife (Mosteller & Tukey, 1977) to estimate the size of standard errors of smoothed and unsmoothed equatings using operationally obtained data and simulated data. Those errors were also compared to standard errors computed by means of the formula given by Lord (1982).

One reason for using this "infinite sample size" equating as a criterion in the simulations is that the standard method used to address unacceptably large uncertainty or error associated with statistical procedures is to increase sample size. Unfortunately, as is well known, the precision so obtained increases in proportion to the square root of the sample size, whereas the scale of a study and, hence, its costs tend to increase at least directly with the sample size. Thus, to increase the precision of a statistical measure by a factor of two (or, equivalently, to reduce the standard error of the estimate of a parameter by half), it is necessary to increase the size of a sample, and thus approximately the cost, by four or more. If the smoothing methods investigated here reduce standard errors by 25 percent, their adoption might be expected to permit the use of samples approximately 56% of the size of current samples, with no loss of accuracy but with a saving in resources expended.

Simulations

The aim of the simulations was to provide data that modeled those which might result from having examinees take ASVAB-like subtests. The range of test lengths investigated covers the range of lengths of subtests in the operational ASVAB. Three test lengths were used — 15 items, 30 items, and 50 items. For each test length, two very similar tests were created in simulation. The tests were not strictly parallel. They were, however, as similar to each other as are ASVAB subtests within a single subject area in ASVAB 8, 9, and 10. (Ree, Mullins, Mathews, & Massey, 1982.) A sample of 2,000 randomly selected simulated examinees was administered one test, while a second sample of 2,000 was administered the other test. (The term "simulee" will be used hereafter to indicate a simulated examinee.) That process was repeated for a total of 100 simulated administrations for each test length. The same two simulated tests were used, but the sample of simulees was drawn anew for each simulated administration. Different simulated samples were used for each of the test lengths. The following paragraphs describe in detail the method of the simulations. Throughout this section, reference is made to random selection and random numbers. The numbers used were generated by a pseudorandom generator, not a totally random generator, as is common in computer simulations. Although determinate, sequences of pseudorandom numbers appear much as random numbers, are indistinguishable from them by most standard tests, and do not repeat the sequence of numbers until millions of numbers have been generated.

Item Response Theory (IRT) (Lord, 1980) is the most explicit, complete, and quantitative theoretical treatment of tests of mental ability. The simulations were therefore carried out within the framework of IRT. Each aspect of the simulated tests and of the simulees was specified in IRT terms in such a way as to model operational subtests in the ASVAB testing program. (See United States Military Entrance Processing Command, 1984, for a description of the ASVAB program.) The exception to this general statement is in connection with the 50-item test. The longest power subtest in the ASVAB is 35 items. The simulated 50-item test was constructed to simulate the same test as it might operate if it were lengthened to 50 items. The test length of 50 items was included to determine the effectiveness of the smoothers with tests of moderate length.

The 15-item test was designed to simulate Paragraph Comprehension, the 30-item test was designed to simulate Arithmetic Reasoning, and the 50-item test was designed to simulate a lengthened version of Word Knowledge. For each of the subtests to be simulated, the statistics which describe the operational subtests were first considered. Subtest reliabilities, classical item statistics, IRT statistics, and means and standard deviations for the operational subtests were obtained from a technical report by Ree et al. (1982). Simulated items were generated at random so that the items' distributions of a, b, and c parameters approximately matched those reported by Ree et al. (1982) for the operational subtests. The resulting tests were then administered in simulation to samples of 2,000 simulees. The resulting item p-values and item-test biserial correlations and such test statistics as mean, standard deviation, skew, kurtosis, and reliability (KR-20) were examined to determine whether or not all of the items were similar to the items which are found on operational ASVAB subtests.

Some of the simulations used to generate the data to evaluate the smoothings were conducted in the same manner as the simulations used to develop the tests. The method used to simulate the administration of a test is similar to a method developed by Ree (1980) for use in a simulation carried out in another context. In order to carry out a simulated administration, a population of examinees was defined with ability normally distributed with a mean ability, θ, of zero and a standard deviation of 1.0. IRT equations allow the computation of the probability that an examinee of some known ability will correctly answer an item with known parameters a, b, and c. The parameter a specifies the steepness of the slope of the item characteristic curve. The parameter b is a measure of the difficulty of an item. The parameter c indicates the probability that an examinee of very low ability will answer correctly. That probability, P, is given by the formula

$$P = c + (1 - c) / (1 + \exp(-1.7\, a\, (\theta - b))).$$

For each of the 2,000 simulees in the sample and for each item, a program computed the probability of an applicant's answering the item correctly. The program then selected a random deviate from a rectangular distribution on the open interval $0 < \theta < 1$. If the deviate was less than the probability of a correct response, then the simulated response was counted as correct; if it was equal or greater, then it was counted as incorrect. Such a simulation results in response vectors which include correct responses due to the joint influences of ability and guessing, just as operational data show both such influences. When all 2,000 simulees had "responded" to all items in a test, the test was scored and analyzed to determine the mean and standard deviation of scores, the item difficulties, the item biserial correlation coefficients, and other statistics.

The resulting test statistics and distributions were compared with the results of the subtests which the simulated tests were designed to match. Items which were too difficult or too easy, that is, items which had *p*-values (or item difficulties) inconsistent with the requirements for ASVAB items, were replaced with items with *a*, *b*, and *c* parameters which would lead to more appropriate *p*-values, and the simulations were rerun. Each of the simulated tests went through several iterations of that process in order to arrive at tests which resembled ASVAB subtests. The refining process was necessary in part because the technical material dealing with the ASVAB does not report the item parameters for the individual items, but gives only summaries. At each test length, the items in one test were chosen to be slightly more difficult than the items in the other test, so that the equatings would not result in virtual identities. Since virtually identical subtests are not found in the ASVAB program, and since they would not require equating if they existed, they were not sought in this project. The technical aspects of the resulting simulated tests did resemble the technical aspects of the ASVAB subtests in every aspect but one. The item-test biserial correlation coefficients in the simulated tests were higher than the corresponding coefficients in the operational subtests.

Although it is not known exactly why the biserials should be higher when the *a*, *b*, and *c* parameters are comparable, three possibilities are evident. The first possibility derives from the relationship between the *a* parameter, or the item discrimination index, and the biserial correlation. Lord (1980, p.33) gives that approximate relation as

$$R_b \cong \frac{a}{\sqrt{1+a^2}}$$

when a is the *a* parameter and R_b is the biserial correlation. The methods used to estimate *a*, *b*, and *c* for the operational items may have overestimated the *a* values (see Ree, 1979, for an evaluation of estimation procedures), with the result that simulations which use the reported *a* values would tend to have higher biserials than would the operational tests. A second and perhaps more likely possibility is that any departure from test theory in the operational setting reduces the resulting biserials. Accidentally mismarked answer sheets, careless errors, omitted items, and many other extraneous response variables can reduce biserials from the theoretically expected value. The simulations were free from such departures from theory. Third, IRT requires the assumption that tests are unidimensional, or that each test measures only one ability. No operational subtest is perfectly unidimensional, but the simulated tests were. It is likely that the perfect unidimensionality of the simulated tests is a significant factor in accounting for the high biserials.

A test characteristic curve was prepared for each test to show the true score, or score which would be found in the absence of measurement error, expected to correspond to each level of ability, designated by theta, from -3 to +3. (See Allen and Yen, 1979, for a discussion of true scores and test characteristic curves.) Similarly, for each test, test information curves were prepared to show test information as a function of ability on the same interval. Figures 1, 2, and 3 show those curves for 15-, 30-, and 50-item tests, respectively.

Technical and statistical details of the tests are presented in Appendix A. Each of the six simulated examinations was "taken" by 100 groups of 2,000 simulees. Either of two methods was used to administer a test in simulation. The first method is that described above in connection with the development of the forms. The second method involved taking a sample of 2,000 observations at random from the Expected Observed Score Distribution (EOSD, described on page 15) for a test. Score distributions were tabulated for each simulated administration. For each test length, 100 equipercentile equatings and smoothings were then performed using the methods described below. The smoothings and equatings were the same for the operational and simulated data and so are described following the description of the operational data.

Figure 1

Test information curves (upper panel) and test characteristic curves (lower panel) for the simulated tests of length 15. The solid lines represent the reference test; the dotted lines represent the experimental test.

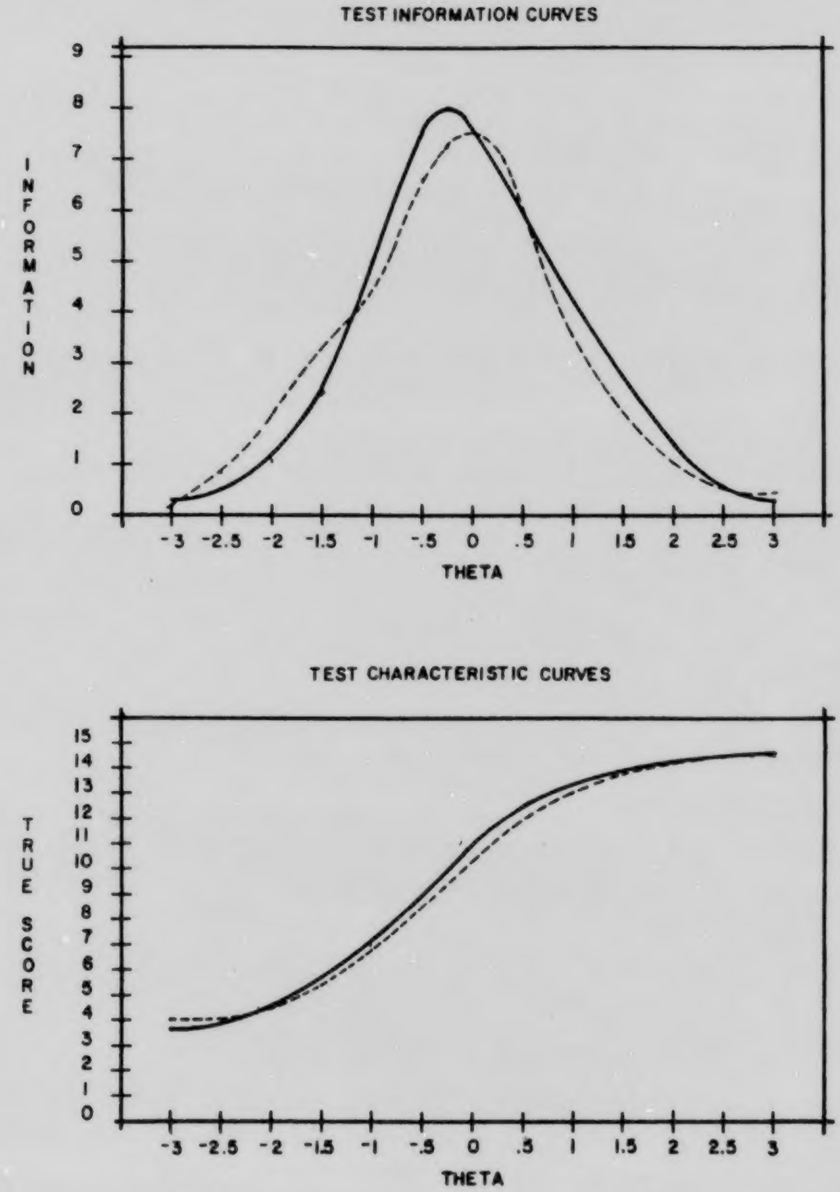


Figure 2

Test information curves (upper panel) and test characteristic curves (lower panel) for the simulated tests of length 30. The solid lines represent the reference test; the dotted lines represent the experimental test.

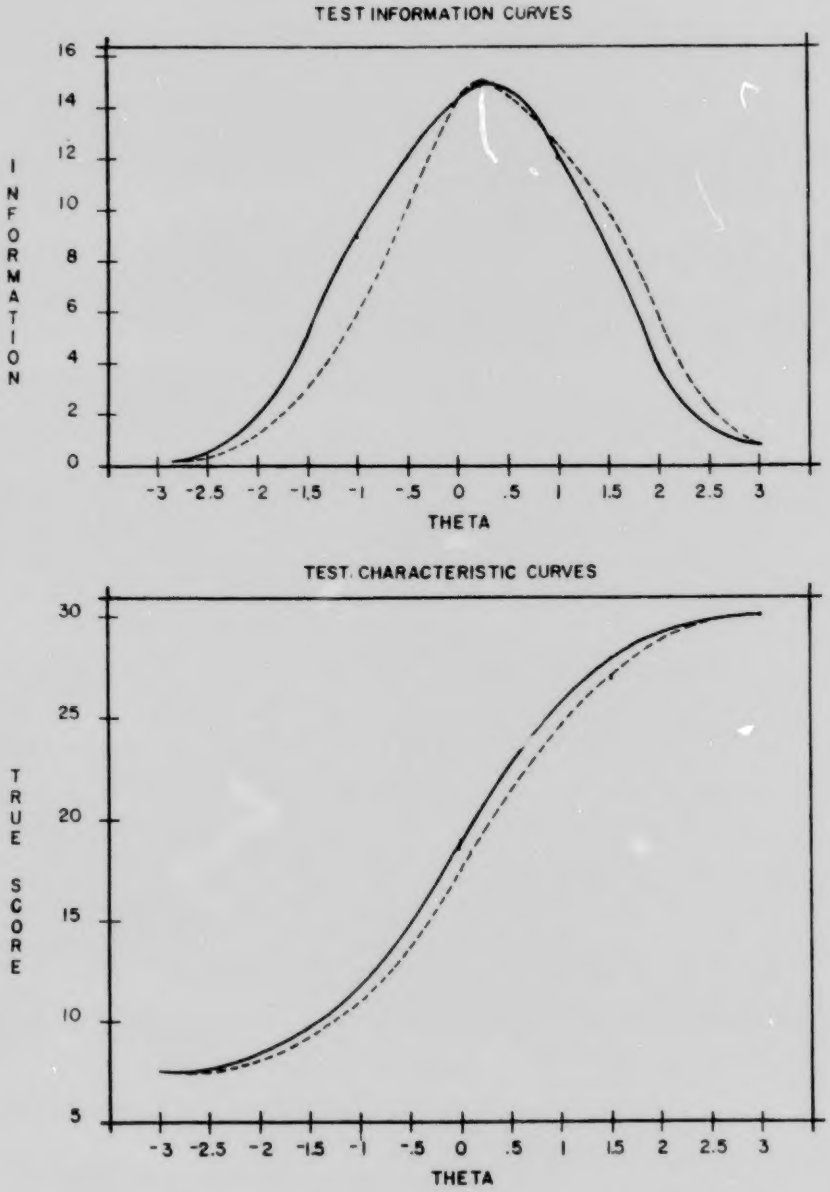
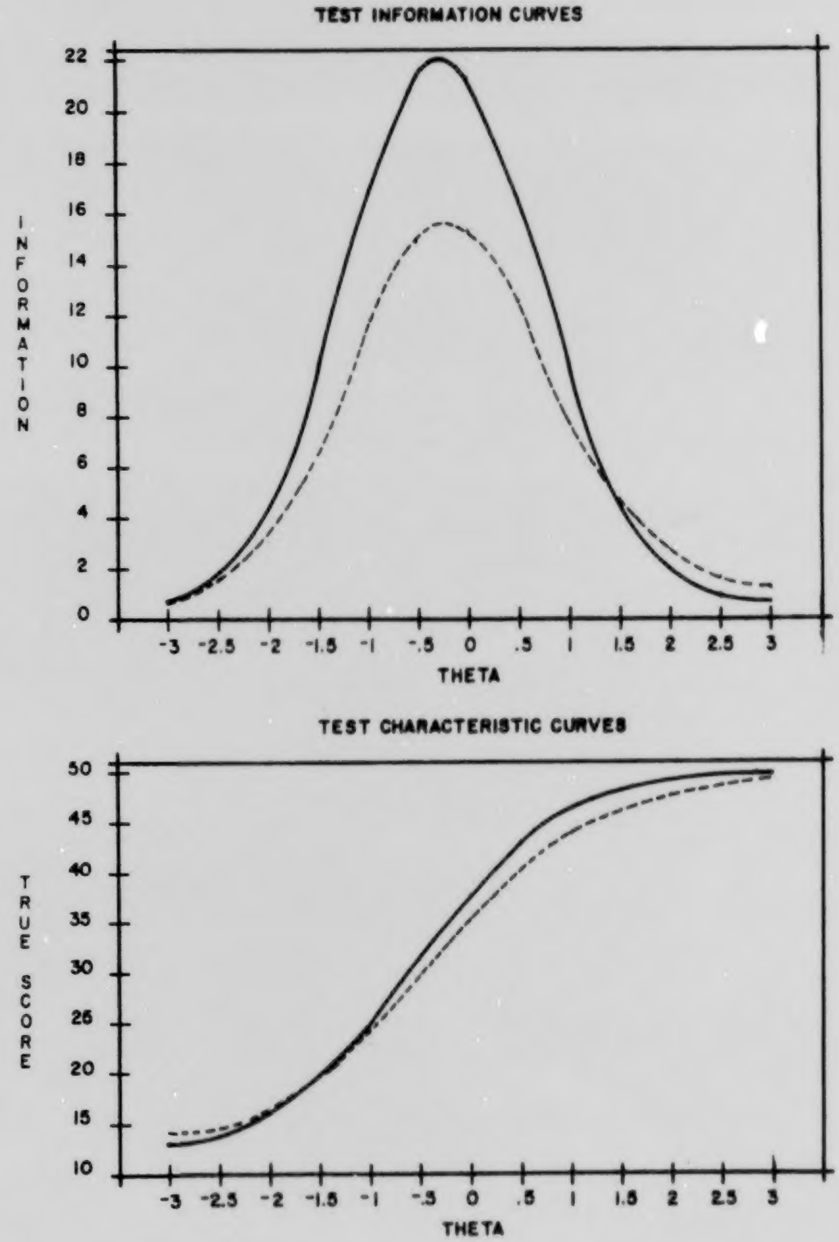


Figure 3

Test information curves (upper panel) and test characteristic curves (lower panel) for the simulated tests of length 50. The solid lines represent the reference test, the dotted lines represent the experimental test.



Criterion Equatings

The preparation of simulated tests allows total control of the simulated test situation. It is, therefore, possible to know in advance the criterion or "true" equating of the tests used. Item response theory makes possible several approaches to the determination of the criterion equating. It is possible, for example, to determine the true scores associated with various abilities or theta values and equate true scores through common theta. Analogously, a variant of true-score equating can be performed, and for each integer true score on the experimental test, the corresponding theta can be computed (usually by means of inverse interpolation); then the score on the reference test which corresponds to that theta can be found. That method has the advantage of giving equated scores for each number-right true score, and interpolation of tabled values is not required. True scores are never known in actuality, however, so that method is not wholly appropriate.

The method used to establish the criterion equatings for the simulations used in the present study is based on the EOSD for each test. An algorithm developed by Lord and Wingersky (1983) was used to prepare distributions of expected observed scores for each of the six simulated tests. In an EOSD, each score has associated with it a proportion of examinees, not a frequency. The distributions model the result of administering the test to an infinitely large number of examinees and observing the relative frequency of each score. The EOSD method of establishing a criterion equating is appropriate because the aim of the present research is to determine methods of smoothing which compensate for the relatively small sample sizes that must be used operationally. By comparing the small sample equatings (N=2,000) with those that result from an "infinite" sample (i.e., those based on the EOSD), the extent of improvement resulting from smoothing is directly observable. The criterion equatings, then, are the unsmoothed equipercentile equatings which result from using the EOSDs in the unsmoothed equipercentile method. Lord and Wingersky (1983) show that such equatings do not differ appreciably from true score equatings based on IRT.

Operational Data

The operational data that were used were taken from a set of ASVAB scores for very large sample sizes (approximately 100,000 examinees) for three roughly parallel forms of each of several subtests. Among those subtests were two forms of Mathematics Knowledge, length of 25 items, and two forms of Electronics Information, length of 20 items. In addition to the frequency distributions of test scores for all examinees, there were available 100 samples of 2,000 scores for each of the four subtests (two forms each of Mathematical Knowledge and Electronics Information). The samples were drawn at random without replacement from the larger samples of 100,000 examinees. Two test lengths were thus available in the operational data: 20 and 25 items. The lengths used were constrained in part by the availability of data and in part by the aim of increasing the generalizability of the study by employing a number of different test lengths for operational and simulated tests.

For the operational data, criterion equatings were established by using the full sample of 100,000 examinees. Although that sample equating is not totally error-free, it is based on a sample 50 times as large as the samples of size 2,000 and so is expected to have sample-dependent errors only approximately one-seventh as large as those found in the small equatings. As with the simulated data, the criterion equatings were unsmoothed equipercentile equatings, as described below. As with the simulated data, 100 reduced sample equatings were made for each of the test pairs, both without smoothing and with each of the 19 smoothing methods.

Equatings

All test equatings were performed using the equipercentile method described by Lindsay and Prichard (1974). For the unsmoothed equatings and the equatings to which only postsmoothing was to be applied, the raw frequency files were equated. When the equatings involved presmoothing, the smoothed frequency estimates were equated. Following the equatings and smoothings (which are described below), each test or simulated test had associated with it a criterion equating, an unsmoothed equating, and 19 smoothed equatings, one for each of the smoothing methods used. The equatings resulted in associations between each observed score on the experimental test and scores on the reference test. Such associations may be expressed by equating tables prepared for operational use by test administors or users. The equating tables, over 2,000 of which were required for this effort, were generated and used by the equating program but were not printed.

Smoothings

The methods used for smoothing the data are listed and described in the introduction. Most of the smoothing methods are sufficiently simple to implement that they require no description beyond that given in the introduction. Two of the postsmoothing methods, however, are more complex and require further description.

The fifth postsmoothing method was the fitting of a logistic ogive to the data. The ogive was fit by the method of the simplex, which is an iterative, rather than an optimal, method. The method requires an initial estimate of the four parameters (upper and lower asymptotes, slope, and location) which define the ogive; it then successively finds better and better sets of points. In this case, "better" implies sets of points with smaller residual sums of squares. The simplex continues to iterate until the values of the four parameters converge to final values. The initial simplex for the first smoothing of the 100 equatings was chosen to be a simplex whose asymptotes were far enough from the points to be smoothed that the part of the ogive passing through the points was an approximation of a straight line. Subsequent smoothings took as their initial simplex the final simplex of the previous smoothing. The method has the disadvantage of being prone to produce solutions which represent local, rather than global, minima. Experience with the method has indicated that it does occasionally fall into such minima; such solutions do not represent the best-fitting functions and may occasionally not fit well at all.

The procedure used here for the fitting of the cubic spline departed in three ways from that used by Kolen (1983). First, Kolen fit two spline functions, one using the equated experimental test scores as the dependent scores, and the other using the reference test scores as the dependent variables. The final equated values were obtained by averaging the equatings resulting from the use of those two spline functions. In order to retain comparability with other smoothing methods used in this research, the experimental test was used as the dependent variable in the fitting of the spline.

The second departure involved the difficulty which was encountered with cubic spline smoothing at lower ends of the score distribution. Kolen (1983), finding similar difficulty at both ends, addressed it by applying the splines only in the interval of test scores ranging from the 5th to the 95th percentile. The shortest of his tests, however, was 40 items, and few examinees scored at either of the extremes. Smoothing by means of cubic splines as described by Reinsch (1967) requires an estimate of the standard errors of the y variables at each duck, but at the lower end point, where frequencies are at or near zero, the standard errors are not defined or do not exist. For the purposes of this investigation, the end standard errors were assigned the value of the closest defined standard error, when "closest" means the numerically closest integer score. The use of a large standard error constrains the spline function to pass through or very close to the point associated with the unsmoothed equating.

Initially, smoothing methods relied heavily on human judgment and experience in passing a line among the points. The hope of those using the more analytic smoothing methods has been that an optimum or nearly optimum method might be found so that judgmental methods would not be necessary. Smoothing could then be automated and thus replicable and objective. The work of Kolen (1983), whose cubic splines have been among the most effective postsmoothing methods described in the literature, has not avoided the necessity of intervening judgment in the application of the smoothing process. For the current project, however, when over 500 applications of the smoothing technique were required, automated smoothing was a necessity. Thus, the third departure was the use of standard errors in the cubic spline fitting procedure.

The smoothers usually resulted in slight changes in the total number of cases in the smoothed distribution as compared to the unsmoothed distribution. The changes were due to the action of the medians or means in lowering unusually high values or raising unusually low values. The total number of cases was always adjusted to the original number of 2,000 by increasing each frequency by whatever proportion was necessary in order that the total frequency equal 2,000. Thus, the shape of the distribution and the relative heights of its frequencies were unchanged by the adjustment. The adjustment resulted in the use of some noninteger frequency values in the equating step, an option permitted by the equating method.

Analysis of Equating Results

Each of the five tests, three simulated and two operational, had associated with it one criterion equating, 100 unsmoothed equatings based on sample sizes of 2,000 (called the "small sample"), and 100 sets of 19 smoothed equatings based on the same samples. The question of interest is the effect of the smoothings on the accuracy of the equatings. The measures used to define the accuracy of the equatings are based on the concept of deviations. A deviation is a difference between an equated score obtained with a small sample and an equated score based on a criterion equating. At each observed (i.e., integer) score on the experimental test, the corresponding score on the reference test was found using the criterion equating. The equated scores were found as decimal fractions not rounded to the nearest integer. The score corresponding to the same experimental test score was then found for the unsmoothed small sample equating and for each of the 19 smoothed equatings. The differences between the equated score based on the criterion equating and the equated score based on the small sample equatings were found for each possible score on the experimental test, for the unsmoothed and for the smoothed equatings, for all 100 replications. These differences, or deviations, were the raw data used for evaluating the smoothings. A deviation, D, associated with a given score on an experimental test, unsmoothed or smoothed by a particular method, is thus defined by the formula:

$$D = x - x'$$

where x equals the equated score based on criterion equating and x' equals the equated score based on small sample equating. Each test thus has as many deviation scores, D, as there are items on a test, plus 1 (for a score of 0). For each of the 100 small sample equatings, the deviations at each score were combined across equatings to give a general measure of deviation at each score. Three such deviation measures were computed.

The first measure is the Root Mean Square Deviation (RMSD), found by taking the square root of the sum of the squares of the deviations across all 100 samples. The second measure is the Average Absolute Deviation (AAD), or simply the mean of the absolute value of the deviations computed across all samples. The third measure is the average of the signed values of the deviations (ASD), found by taking the mean of the deviations across all 100 replications. ASD differs from AAD in that the absolute values are not found before the mean is computed. ASD is sometimes called "bias," or "statistical bias," but in the context of testing the term "bias" denotes other phenomena and so is less appropriate than "ASD." Positive values of ASD indicate that the small sample equating resulted in a value which was generally lower than the criterion equating values, whereas negative values indicate the opposite. These three measures, RMSD, AAD, and ASD, were found for each score point on each test for the unsmoothed and for each of the 19 smoothed equatings, across all 100 sample equatings.

The three measures of deviation taken together allow an evaluation of the effects of the smoothing methods. The AAD and RMSD both give numbers which represent the unsigned magnitude of an average deviation. The AAD is a straight arithmetic mean of absolute values, while the RMSD has the effect of weighting (or emphasizing) the deviations which are far from the criterion equating. The computation of RMSD is similar to the computation of the standard deviation, which is also sometimes called root mean square deviation from the mean. The difference between the AAD and the RMSD is an indication of the extent to which the distribution has outlying values. If the RMSD is considerably larger than the AAD, then a large number of outliers is suspected. The ASD measure averages the deviations as does AAD, but it includes their sign. The resulting ASD shows how far the mean of the equated values for all 100 samples is above or below the value given by the criterion equating. This is a significant value for two reasons. First, equipercentile test equating has not been shown to be statistically unbiased; the ASD estimates how large the ASD (or bias) actually is. Second, methods which reduce RMSD or AAD may increase ASD. Thus, one must consider RMSD, AAD, and ASD in evaluating a smoothing technique.

Standard Errors

Two cross-checks were made to ensure the accuracy of the methods used to determine the RMSD. First, the standard errors of equipercentile equating were determined using a formula derived by Lord (1982). The resulting standard errors, one at each score level which was at or above chance, or expected guessing score level, were compared with the RMSDs obtained from the simulated test administrations, from the operational data, and from the results of the jackknifing. The observed RMSD values should be empirical estimates of the same standard errors which the Lord (1982) standard errors represent. In each case, the data from the criterion equating, with sample size assigned to be 2,000, were used to develop the standard errors.

The method of Lord (1982) allows calculation of standard errors for the unsmoothed case of equipercentile equating. The simulations discussed above allow empirical estimations of the standard error for all of the smoothing conditions as well as for the unsmoothed condition. The agreement or nonagreement of the Lord formula values with the values generated through simulation indicate the extent to which the simulations are appropriate in the unsmoothed case. There is no corresponding analytical cross-check on the values of the standard errors in the smoothed cases because formulae for standard errors in those cases do not exist. In order to provide corroborating or noncorroborating estimates of the standard errors for smoothed equatings, the equatings were conducted with the use of Tukey's statistical jackknife.

The jackknife (Mosteller & Tukey, 1977) provides an estimate of the standard error of a procedure regardless of whether or not analytical formulae for such errors are available.

Briefly, the jackknife requires dividing a sample into a number of subsamples, then performing the analysis in question (smoothed equipercentile test equating in this case) once with each subsample deleted. In this case, samples of size 2,000 were divided at random into 40 groups of size 50. The smoothings and equatings were performed 40 times for each test length, once without each of the subsamples. The results of 40 equatings were combined according to the procedures of jackknifing in order to obtain estimates of the standard errors of interest. Estimated standard errors were computed for each of the nonchance score levels on the tests; those standard errors were averaged over all such test scores. Thus, each test combined with each smoothing method resulted in a mean standard error of equating as estimated by the jackknife and as estimated by the small sample equatings.

The samples of size 2,000 used for jackknifing were based on the best available estimates of the shape of the observed score distribution in the large sample case. For the simulated tests, the 2,000 cases were assigned scores based on the proportions of scores developed for the expected observed score distribution. For the operational tests, the 2,000 cases were assigned scores based on the proportions observed in the largest samples which the monitoring agency provided. The same proportions were used for the development of the criterion equatings when using the simulations to develop the deviation measures.

III. RESULTS

Graphic presentation of the results of this effort are presented in Appendix B, Figures B - 1 through B - 95. Each figure is for one test length and one method of smoothing. The figures are grouped so that the effects of each smoothing method can be evaluated across all five tests (three simulated and two operational). The results of the simulations are presented first, with the test lengths of 15, 30, and 50. Following those are the two operational tests, of length 20 and 25. Each figure presents three panels. Each panel shows measures of deviation as a function of the raw score on the experimental test, both with and without smoothing. In each figure, the top panel shows the effect of smoothing on RMSD, the middle panel shows its effect on AAD, and the bottom panel shows the effect on ASD.

Two functions are shown on each panel of each figure. The continuous line shows the RMSD, AAD, or ASD which results from equating samples of size 2,000 without smoothing, while the + characters indicate the RMSD, AAD, or ASD when the same samples are equated with smoothing. Each point on the graph is an average of 100 deviations, or differences between the criterion equating and the small sample equating. The vertical axis of each graph has been scaled so that the maximum measure of deviation takes up a large portion of the axis. In comparing different figures, one should note the magnitude of the axes. When there is a horizontal line in a graph other than the top or bottom line of the graph, that line represents zero deviation. For the figure panels which depict AAD and RMSD, + signs which lie below the continuous line indicate that an improvement, or a reduction of deviations, resulted from smoothing. The situation with ASD is slightly more complex, since ASD may be either positive or negative. Improvement, or reduction of ASD, is indicated when the + signs lie either between the continuous line and the x-axis of the graph, or closer to the x-axis than the continuous line.

It was found that with some smoothing methods, especially the presmoothing methods, smoothing resulted in large increases in the deviation measures for very low test scores. In some cases the increases were so great that graphing them required such a large rescaling of the figures that the more important deviations in the middle ranges of the test could not be represented. These large induced deviations are seen as being of little interest because they occurred at score values which were lower than the guessing level on a test, and so were not associated with meaningful measures of ability. In order to show the more relevant deviations effectively, the figures do not present information on the levels of RMSD, AAD, or ASD at test scores below the guessing level for each test.

The graphs may be considered in a number of ways. It is suggested that in considering the graphs, particular attention be given to the top panels, where the effect of smoothing on RMSD is shown. If a particular smoother is effective in reducing RMSD across all five test lengths, then it should be considered further. In particular, its effects on ASD should be considered, in order to determine whether it is effective in reducing ASD or, as frequently happens, whether it increased ASD. The figures are summarized in tables presented later in this section, but perusal of the figures gives more detail concerning the effects of the smoothers.

In considering the figures, it is helpful to note that if ASD is near zero, then values of RMSD and AAD which are nonzero result from deviations which are distributed approximately evenly between positive and negative deviations. If, in contrast, the ASD is not zero, then the deviations are predominantly above or below zero. That is shown particularly clearly in Figure B - 32, which shows the result of presmoothing by the method of the negative hypergeometric. The unsmoothed equating (solid line) and smoothed equating (+ signs) at the lower end of the scores show that the smoothing increased RMSD and AAD moderately. The ASD, however, increased from near zero to about 0.15 point. The interpretation of this and similar effects in other figures is that the deviations increased moderately in their size, and became predominantly positive in sign, reflecting an increase in local bias.

Local effects, such as the preceding, make it difficult to summarize the effects of the smoothers in tables without obscuring important effects. It is suggested that the figures present the results of this study more effectively than can the summary tables and, hence, should be consulted not only in order to obtain general impressions of the effects of the smoothers, but also to verify impressions obtained from the summary tables.

Tables 1 through 20 summarize the information shown in Figures B - 1 through B - 95. Each table corresponds to one smoother. Table 1 shows the RMSD, AAD, and ASD as averaged across all test scores above the guessing level with no smoothing. The averages of the ASD were taken over the absolute values of ASD so that positive and negative values would not cancel out. The subsequent tables are associated with the smoothed equatings. The averages of RMSD, AAD, and ASD are presented as proportions of the deviations in Table 1. Thus, figures less than 1 indicate that smoothing reduced the deviation, while figures greater than 1 indicate an increase in deviations. For example, a figure of 0.2 indicates that a particular smoothing method reduced the mean measure of a deviation to 20% of its unsmoothed value, when that mean is taken over all scores on a test which are above the chance level.

These tables indicate the effects of the smoothers in a global sense. The effects are averaged over all scores above chance and so may obliterate the locally high deviations. The standard errors of equating and the related measures of RMSD, AAD, and ASD are summary measures of the extent to which a test equating is subject to sample-dependent error. Procedures which result in the reduction of such measures increase the merit of equating methods.

Jackknifing

Table 21 compares the standard errors of unsmoothed equating as estimated by Lord's analytic formula (1982) with those estimated by means of repeated reduced sample equating (i.e., RMSD from simulated or operational tests) and those estimated by means of the jackknife. The standard errors are presented in the metric of test items. They are averaged over all test scores which are higher than chance level. The standard errors thus indicate that standard errors of equating vary with the length of the tests, and vary with the method used to estimate them.

Table 22 presents the standard errors of smoothed equatings, as estimated by the RMSD of the reduced sample equatings (100 samples of 2,000) and as estimated by Tukey's jackknife (Mosteller & Tukey, 1977). In order to facilitate evaluating the effects of smoothing, the standard errors of the smoothed equatings are presented as proportions of the unsmoothed equatings given in Table 21. Thus, values greater than 1.00 indicate that smoothing increased the standard errors, whereas values less than 1.00 indicate a reduction in standard error.

Table 1. Mean Measures of Deviations for Unsmoothed Equatings

	Mean Deviation		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.134	.106	.009
30	.269	.214	.016
50	.439	.348	.028
	Operational Tests		
20	.184	.145	.029
25	.242	.192	.015

Note. Tabled values represent RMSD, AAD, ASD averaged over all samples at all scores above chance level.

Table 2. Summary of the Averaged Effects of Presmoothing by the Method of 3-Point Moving Medians

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	1.004	1.013	1.523
30	1.002	1.006	1.034
50	1.058	1.052	.821
	Operational Tests		
20	.997	.996	1.003
25	.994	.999	1.233
Mean	1.011	1.013	1.123

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 3. Summary of the Averaged Effects of Presmoothing by the Method of 5-Point Moving Medians

Test Length	Proportion of Mean Deviations RMSD	AAD	ASD
	Simulated Tests		
15	1.021	1.046	1.841
30	.993	.994	1.352
50	1.027	1.027	1.362
	Operational Tests		
20	1.016	1.028	.881
25	1.007	1.024	1.749
Mean	1.013	1.024	1.473

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 4. Summary of the Averaged Effects of Presmoothing by the Method of 3-Point Moving Weighted Averages

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.962	.963	1.183
30	.974	.975	.846
50	.979	.981	1.303
	Operational Tests		
20	.953	.948	1.100
25	.969	.970	1.606
Mean	.967	.968	1.208

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 5. Summary of the Averaged Effects of Presmoothing by the Method of 5-Point Moving Weighted Averages

Test Length	Proportion of Mean Deviations RMSD	AAD	ASD
	Simulated Tests		
15	.994	.995	1.458
30	.990	.992	.777
50	.990	.992	.946
	Operational Tests		
20	.988	.986	1.153
25	.985	.987	1.187
Mean	.989	.990	1.104

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 6. Summary of the Averaged Effects of Presmoothing by the Method of 5-Point Moving Weighted Averages with Root Transformation

Test Length	Proportion of Mean Deviations		
	RMSD	AAD	ASD
		Simulated Tests	
15	.995	.996	1.470
30	.990	.993	.801
50	1.002	1.006	1.040
		Operational Tests	
20	.986	.982	1.129
25	.984	.988	1.248
Mean	.991	.993	1.138

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 7. Summary of the Averaged Effects of Presmoothing by the Method of 4253H Twice

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	1.019	1.036	2.721
30	1.013	1.013	1.364
50	1.034	1.036	.920
	Operational Tests		
20	.980	.981	1.180
25	.992	.992	1.526
Mean	1.007	1.012	1.542

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 8. Summary of the Averaged Effects of Presmoothing by the Method of Negative Hypergeometric

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.891	.903	2.919
30	.865	.867	3.596
50	.852	.861	3.453
	Operational Tests		
20	.905	.908	2.008
25	.966	.989	7.479
Mean	.896	.906	3.891

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 9. Summary of the Averaged Effects of Postsmoothing by the Method of Linear Regression

Test Length	RMSD	AAD	ASD
	Proportion of Mean Deviations		
	Simulated Tests		
15	.969	.995	7.181
30	1.131	1.192	10.905
50	1.346	1.385	13.211
	Operational Tests		
20	.917	.941	2.684
25	1.243	1.300	13.639
Mean	1.121	1.163	9.524

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 10. Summary of the Averaged Effects of Postsmoothing by the Method of Quadratic Regression

Test Length	Proportion of Mean Deviations RMSD	 AAD	 ASD
	Simulated Tests		
15	.992	1.001	2.713
30	1.031	1.042	5.362
50	1.754	1.942	18.896
	Operational Tests		
20	.971	.993	1.394
25	.960	.966	4.249
Mean	1.141	1.189	6.523

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 11. Summary of the Averaged Effects of Postsmoothing by the Method of Cubic Regression

Test Length	Proportion of Mean Deviations		
	RMSD	AAD	ASD
	Simulated Tests		
15	1.077	1.062	2.626
30	1.054	1.053	3.131
50	1.318	1.252	3.333
	Operational Tests		
20	1.118	1.130	2.048
25	.996	1.000	2.199
Mean	1.113	1.100	2.667

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 12. Summary of the Averaged Effects of Postsmoothing by the Method of Orthogonal Regression

Test Length	Proportion of Mean Deviations RMSD	AAD	ASD
	Simulated Tests		
15	.876	.890	6.779
30	1.012	1.056	9.948
50	.962	1.001	9.103
	Operational Tests		
20	.884	.907	2.708
25	1.175	1.234	13.048
Mean	.982	1.018	8.317

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 13. Summary of the Averaged Effects of Postsmoothing by the Method of Logistic Ogive

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.872	.886	6.767
30	.970	1.003	8.883
50	.940	.979	8.980
	Operational Tests		
20	.879	.902	2.651
25	1.170	1.230	13.001
Mean	.966	1.000	8.056

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 14. Summary of the Averaged Effects of Postsmoothing by the Method of Cubic Splines

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.914	.917	1.548
30	.935	.927	2.086
50	.984	.984	1.773
	Operational Tests		
20	.935	.932	.956
25	.928	.927	1.364
Mean	.939	.937	1.545

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 15. Summary of the Averaged Effects of Postsmoothing by the Method of 5-Point Moving Weighted Averages

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.984	.985	1.115
30	.990	.989	.980
50	.994	.993	1.013
	Operational Tests		
20	.985	.985	.995
25	.990	.990	1.069
Mean	.989	.989	1.035

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 16. Summary of the Averaged Effects of Combined Smoothing by the Method of Combined Presmoothing by Negative Hypergeometric and Postsmoothing by Orthogonal Regression

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.831	.841	6.034
30	.965	1.011	9.995
50	.957	1.018	8.903
	Operational Tests		
20	.771	.780	2.078
25	1.064	1.121	11.742
Mean	.918	.954	7.750

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 17. Summary of the Averaged Effects of Combined Smoothing by the Method of Combined Presmoothing by Negative Hypergeometric and Postsmoothing by Quadratic Regression

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.898	.907	2.504
30	.916	.923	6.312
50	1.143	1.227	9.310
	Operational Tests		
20	.397	.155	.246
25	.885	.898	5.855
Mean	.848	.822	4.845

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 18. Summary of the Averaged Effects of Combined Smoothing by the Method of Combined Presmoothing by Negative Hypergeometric and Postsmoothing by 5-Point Moving Weighted Averages

Test Length	Proportion of Mean Deviations		
	RMSD	AAD	ASD
	Simulated Tests		
15	.890	.904	2.932
30	.866	.868	3.614
50	.852	.861	3.466
	Operational Tests		
20	.904	.907	2.012
25	.966	.989	7.479
Mean	.896	.906	3.901

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 19. Summary of the Averaged Effects of Combined Smoothing by the Method of Combined Presmoothing by 3-Point Moving Weighted Averages and Postsmoothing by 5-Point Moving Weighted Averages

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.958	.959	1.285
30	.970	.971	.901
50	.976	.978	1.292
	Operational Tests		
20	.948	.944	1.106
25	.965	.965	1.601
Mean	.963	.963	1.237

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 20. Summary of the Averaged Effects of Combined Presmoothing by Negative Hypergeometric and Postsmoothing by Cubic Splines

	Proportion of Mean Deviations		
Test Length	RMSD	AAD	ASD
	Simulated Tests		
15	.885	.898	2.968
30	.878	.875	3.985
50	.855	.863	3.613
	Operational Tests		
20	.902	.905	1.998
25	.966	.990	7.463
Mean	.897	.906	4.005

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing, as presented in Table 1. Averages taken over all samples at all scores above chance level.

Table 21. Standard Errors of Unsmoothed Equating Estimated by Three Methods

Kind of Test.	Simulated			Operational	
Length of Test.	15	30	50	20	25
Method of Estimation					
Lord's Formula	0.15	0.30	0.51	0.18	0.25
Average of 100 Samples	0.13	0.27	0.44	0.18	0.24
Jackknifing	0.15	0.25	0.47	0.17	0.23

Note. Standard errors were averaged over all scores above the chance or guessing level.

Table 22. Proportional Change in Standard Errors as a Result of 19 Smoothing Methods

Kind of Test:	Simulated			Operational	
Length of Test:	15	30	50	20	25
			Presmoothing		
Method of Estimation					
			3-point moving median		
RMSD 100 Samples	1.00	1.00	1.06	1.00	.99
Jackknifing	1.07	1.00	1.02	1.05	1.02
			5-point moving median		
RMSD 100 Samples	1.02	.99	1.027	1.02	1.01
Jackknifing	1.16	1.01	1.00	1.00	1.02
			3-point moving weighted averages		
RMSD 100 Samples	.96	.97	.98	.95	.97
Jackknifing	.98	.98	.98	.94	.98
			5-point moving weighted averages		
RMSD 100 Samples	.99	.99	.99	.99	.99
Jackknifing	1.01	1.00	.99	.98	1.00
			5-point moving weighted averages with root transformation		
RMSD 100 Samples	1.00	.99	1.00	.99	.98
Jackknifing	1.01	1.00	1.00	.98	1.01
			4253H Twice		
RMSD 100 Samples	1.02	1.01	1.03	.98	.99
Jackknifing	1.09	.99	.97	.98	.96
			Negative hypergeometric		
RMSD 100 Samples	.89	.87	.85	.91	.97
Jackknifing	.89	.80	.88	.81	.87

Table 22, continued

Kind of Test:	Simulated			Operational	
Length of Test:	15	30	50	20	25
Postsmoothing					
linear regression					
RMSD 100 Samples	.97	1.13	1.13	.92	1.24
Jackknifing	.76	.82	.93	1.08	.88
quadratic regression					
RMSD 100 Samples	.99	1.03	1.75	.97	.96
Jackknifing	.97	.92	.99	.99	.86
cubic regression					
RMSD 100 Samples	1.08	1.05	1.32	1.12	.99
Jackknifing	.99	1.03	1.45	1.19	.93
orthogonal regression					
RMSD 100 Samples	.87	1.01	.96	.88	1.18
Jackknifing	.70	.85	2.15	1.09	.89
logistic ogive					
RMSD 100 Samples	.87	.97	.94	.88	1.17
Jackknifing	.70	.85	2.03	1.09	.88
cubic splines					
RMSD 100 Samples	.91	.94	.98	.94	.93
Jackknifing	1.00	1.01	1.01	.99	.93
5-point moving weighted averages					
RMSD 100 Samples	.98	.99	.99	.99	.99
Jackknifing	.95	.99	.99	.98	.99

Table 22, continued

Kind of Test:	Simulated			Operational	
Length of Test:	15	30	50	20	25
Combined Smoothers					
negative hypergeometric + orthogonal regression					
RMSD 100 Samples	.83	.97	.96	.77	1.06
Jackknifing	.74	.71	.78	.65	.77
negative hypergeometric + quadratic regression					
RMSD 100 Samples	.89	.91	1.14	.40	.89
Jackknifing	.91	.78	1.01	.86	.85
negative hypergeometric + 5-point moving weighted averages					
RMSD 100 Samples	.89	.87	.85	.90	.97
Jackknifing	.89	.80	.87	.81	.87
3 point moving weighted averages + 5-point moving weighted averages					
RMSD 100 Samples	.96	.97	.98	.95	.97
Jackknifing	.97	.97	.98	.93	.97
negative hypergeometric + cubic splines					
RMSD 100 Samples	.89	.88	.86	.90	.97
Jackknifing	.89	.80	.88	.81	.87

Note. Table entries show the magnitude of standard error estimates for smoothed equatings when such standard errors expressed as proportions of the corresponding unsmoothed equatings.

IV. DISCUSSION

This section considers first the presmoothing methods, then the postsmoothing methods, then the combined smoothers. It then presents conclusions based on the results, followed by a discussion of the limitations of the study. Finally, recommendations for operational implementation and for further study are presented.

To evaluate the effects of smoothing, particularly its effects on deviations, it is helpful to consider such deviations within the context of the accuracy of ability or achievement tests more generally. The standard errors of equating discussed in this report are not the only measurement errors which arise in the testing process. There are also standard errors of measurement that are intrinsic to any test which is not perfectly reliable. The following formula (Allen & Yen, 1979) relates reliability (R), standard error of measurement (SE), and test score standard deviation (SD).

$$SE = SD * \sqrt{1 - R} \ .$$

Thus, the standard error of measurement for the experimental test of length 15, based on a reliability (KR-20) estimate of .80 and a standard deviation of 3.28, both given in Table A-1, is 1.47. Based on data from the same table, the standard error of measurement for the experimental test of length 30 is 2.20, and that for the experimental test of length 50 is 2.74. The corresponding average standard errors of equating, given in Table 21, as estimated by Lord's formula, are .15, .30, and .51. Thus the standard error of equating ranges from approximately only 10 to 20 percent of the standard error of measurement.

Presmoothing

Consideration of Figures B-1 through B-5 shows that smoothing by the method of 3-point moving medians had no consistent beneficial effect. Frequently it resulted in less accurate equatings than unsmoothed equatings. These effects are summarized in Table 2. The means of the deviation measures show that, on the average, the smoother was harmful.

Similar results are obtained from the use of 5-point moving medians (Figures B-6 to B-10 and Table 3). There is no consistent beneficial effect and frequent deleterious effects on all three measures of deviation. Whatever local gains are achieved are offset by losses elsewhere.

The method of 3-point moving weighted averages, whose results appear in Figures B-11 to B-15 and Table 4, is the first method to show generally encouraging results, although the gains are modest. The gains are particularly evident on the 15-item simulated test and the 20-item operational test. There is a modest increase in the ASD at the high score levels in both tests.

The method of 5-point moving weighted averages (Figures B-16 to B-20 and Table 5) has generally negligible effects on all three measures of deviation.

The result of applying the method of 5-point moving weighted averages with root transformation, as shown in Figures B-21 to B-25 and Table 6, is virtually identical to the result of applying the method of 5-point moving weighted averages without root transformation, as described above. There was no significant benefit achieved.

Smoothing by the method of 4253H Twice (Figures B-26 to B-30) was generally ineffective and resulted in local increases and local decreases in the measures of deviation. Table 7 bears out that impression by showing harmful or minimal effects.

Finally, the results of smoothing by means of the negative hypergeometric (Figures B-31 to B-35) are the first to show consistent improvement in RMSD and AAD as a consequence of smoothing. The effects are particularly impressive with the simulated tests, presumably in part because the criterion equatings for those tests are nearly perfect, not estimated from very large samples. The gains are not uniform across the tests. On the shorter tests at lower scores, the measures of RMSD and AAD actually increase as a consequence of using the negative hypergeometric. The considerable decreases in RMSD and AAD are also indicated in Table 8. The beneficial effects of the negative hypergeomtric do not extend to the measures of ASD. The ASD increases both globally and locally, sometimes quite dramatically. These increases were expected at the lower end of the test, where guessing is a factor, but increases at the upper end were not expected. It must be noted, however, that as Table 1 shows, the ASD figures were low initially, so that a tripling of ASD may still denote an acceptably low level. The question of what amount of ASD may be acceptable is complex. Until there are equating methods which can be shown to be consistent, sufficient, efficient, and unbiased, it will be necessary to balance such properties against each other to determine the mix which is optimal for a given purpose. The largest increase in ASD occurred for the test length 50 (Table 8). The increase, by a factor of approximately 7.5, resulted in an increase in the mean ASD (Table 1) from 0.015 score points to 0.11 score points. The mean RMSD for the same test was 0.24 without smoothing, and 0.23 with smoothing. Thus for the 50-item test used in this study the increase in ASD was greater than the reduction in RMSD, although the resulting ASD was only half the magnitude of the RMSD.

If an increase in ASD is less than the decrease in RMSD, then the net benefit may make the use of a smoother which increases ASD justifiable. An increase in ASD may be more acceptable when two tests are equated so that they may be used interchangeably than the same increase would be when the objective of the equating is to replace one operational test with another. If two tests are used interchangeably, then a systematic tendency to deviations in one direction on one test will be offset by scores on the other test. Thus, if the forms of the test are administered at random to examinees, there will be no expected advantage to any examinee. If, in contrast, a test is equated to another so that the older test may be replaced, then ASD will result in equated scores which give results which differ systematically from the scores expected on the test which was replaced.

Why is it that the negative hypergeometric smoothing method outperforms the other presmoothers? It is likely that it is in part because that smoother takes into account all of the information in a distribution's mean and standard deviation in arriving at the smoothed frequency for each point. The other presmoothers respond only to local conditions and so may incorporate, rather than eliminate, some sample-dependent local fluctuations. Although the negative hypergeometric does require the assumption that all items are equally difficult, an assumption usually contradicted in practice, its success as a presmoother indicates that its use is robust against violation of this assumption. Furthermore, among the seven presmoothers investigated, only the negative hypergeometric is based on a mathematical model of testing. The other smoothers work by applying general algorithms which have been shown to be useful in a wide variety of circumstances. It appears that those smoothers do not bring the sample score distributions closer to the shape of the distribution of parent population, whereas the negative hypergeometric does. The negative hypergeometric does so, however, at the cost of increased ASD at some specific test scores.

Postsmoothers

The use of linear regression postsmoothing, as shown in Figures B-36 to B-40, resulted in modest reductions of RMSD and AAD at the middle ranges, but increases at the upper score ranges. The increases in the deviations of the upper score ranges are especially prominent in the simulated test of length 50 (Figure B-38). Since that pattern turns up also with the quadratic and cubic regression postsmoothers (Figures B-43 and B-48), it merits consideration. First, those deviations were not due to the results of the monotonicity constraint imposed on the curvilinear regression smoothing. Although the smoothing algorithm contained the provision for the use of that constraint where needed, it was in fact never required for the data analyzed for this effort. The deviations in the case of the unsmoothed equatings are modest, whereas the smoothed deviations are considerably greater for scores above 47. The concomitant increases in ASD indicate that the increases in RMSD and AAD are due not necessarily to greater variability but rather to consistent deviations in one direction. Furthermore, between tests, the departures are sometimes in one direction, sometimes in the other, as the contrast of the right end of the ASD panels in Figures B-36 and B-37 shows. That this pattern of deviations occurs not only in the curvilinear smoothings, but also in the linear regression smoothings further confirms that it is not due to any nonmonotonicity of the curvilinear smoothing functions, but to the inability of the functions to follow the points adequately. Table 9 summarizes the effects of the linear regression smoother.

Figures B-41 to B-45 present the effects of postsmoothing with quadratic equations. They indicate modest benefits locally. Improvements in RMSD and AAD are partially offset by increases in ASD. Again, deviation measures tend to be high at the upper end, especially with the 50-item test. Table 10 shows the nearly 20-fold increases, from 0.028 to 0.53, in ASD for the 50-item test.

Use of cubic polynomial regression smoothing has less benefit than does quadratic regression in most cases (Figures B-46 through B-50), but it also causes less increase in RMSD at high scores, and less of an effect on ASD. Since a cubic function can follow a given curve more accurately than can a quadratic function, one would expect that the cubic regression smoothing would lead to more accurate equating than linear or quadratic regression smoothing. Findings to the contrary suggest that the cubic functions may have been following and fitting sample-dependent fluctuations in the individual equatings. Table 11 shows its general effectiveness.

Smoothing by means of orthogonal regression (Figures B-51 to B-55) had effects which were very similar to those which resulted from the use of linear regression. The deleterious effects at the high end of the test, however, were less pronounced. Again, there were considerable increases in ASD, although the direction of those increases was not consistent. As Table 12 shows, orthogonal regression was especially variable in its effects on RMSD.

Postsmoothing by means of the logistic ogive, the results of which are shown in Figures B-56 to B-60 and summarized in Table 13, resulted in modest reductions in RMSD and AAD, at the usual cost of increases in ASD, and with the previously noted problems at the highest scores. On the whole, the results of smoothing with the logistic ogive are modestly encouraging.

The results of smoothing by cubic smoothing splines (Figures B-61 to B-65 and Table 14), are the most promising of the results using postsmoothing methods. There are modest reductions throughout in the amounts of RMSD and AAD, combined with no end point problems at either end, and no particularly severe problem with increases in ASD. Table 14 shows that the gains, though modest, are consistent for the cubic smoothing spline method.

Finally, as Figures B-66 through B-70 show, the effect of postsmoothing by 5-point moving weighted averages was virtually nil at all scores and with all three measures. Table 15 confirms the general lack of effect.

Combined Smoothers

Combining negative hypergeometric presmoothing with postsmoothing by means of orthogonal regression (Figures B-71 to B-75) has effects on RMSD and ASD which are greatly inferior to the effects of negative hypergeometric presmoothing alone at the higher score points. This implies that the fitted orthogonal regression smoothing line does not follow the equipercentile points effectively.

The same problems are evident, though to a lesser degree, with the combination of the negative hypergeometric presmoothing and quadratic regression postsmoothing (Figures B-76 to B-80). Evidently the curvilinear function can follow the equating points better than can the straight line. Increases at the higher scores on the 50-item test are evident (Figure B-78), but, in contrast, the combination was very successful for the 20-item test (Figure B-79). The latter, in fact, was the most effective combination seen in this study for any of the five tests considered.

The combination of presmoothing by means of the negative hypergeometric and postsmoothing by the 5-point moving weighted averages (Figures B-81 through B-85) again does not result in gains beyond those achieved with the negative hypergeometric presmoothing alone. The difference in scales makes that difficult to perceive, but it is confirmed by Tables 8 and 18, which have entries that are almost equal to each other.

Figures B-86 through B-90 show that combining the presmoothing method of 3-point moving weighted averages with the postsmoothing method of 5-point moving weighted averages results in slight reductions in RMSD in some cases (Figure B-86), and negligible increases in ASD. Table 19 shows that although the gains are slight, they are consistent.

Finally, Figures B-91 through B-95 show the effects of combining presmoothing by the negative hypergeometric with postsmoothing by cubic smoothing splines. RMSD is decreased at all test lengths, as is AAD, whereas ASD is generally increased, especially with the test of length 25. The effects on ASD vary particularly strongly as a function of test score, as Figure B-92 shows well.

Jackknifed Estimates

The close agreement of the standard errors as estimated by the three methods, shown in Table 21, supports the contention that each of the three estimation methods is both appropriate and correctly executed. Although there are slight differences in the estimates, they are not large enough to call into question the appropriateness of the methods.

The similarity of results continues, for the most part, in Table 22, where the results of smoothing are summarized for all 19 smoothing methods as estimated by RMSD averaged over 100 reduced sample equatings and by jackknifing. There are, however, some cases of disagreement, such as that in the case of postsmoothing by means of linear regression. There, jackknifed results indicate that the method is more generally effective than do the results of the averaged RMSD figures. The jackknifing was applied to groups of 2,000 which were based on population expected proportions, not on small samples. As a result, it is plausible to suggest that the straight line fit the jackknifed samples better than the "true" samples of 2,000 used in the computation of RMSD because the population equating deviates from a straight line less than the small samples do. A similar possibility is evident for the case of combined negative hypergeometric and logistic ogive smoothing.

In general it is difficult to say that either estimation method is unequivocally better than the other, but since the jackknife used data based on population proportions, it is to be expected that analytic functions should fit such data better than they would fit data with sample-dependent fluctuations. In any case, the jackknife does not allow estimations of ASD or AAD.

Since the two methods do give somewhat divergent results in some cases, a conservative criterion for the recommendation of adopting a smoothing method is that the method should appear advantageous with both estimation techniques, mean RMSD and jackknifing. The only method meeting that criterion at all test lengths is the method of presmoothing by the negative hypergeometric.

Conclusions

One presmoother and one postsmoother stand out as deserving further study and consideration for future operational use. The presmoother is the negative hypergeometric; the postsmoother is the cubic smoothing spline.

When its effect is estimated by jackknifing, the cubic smoothing spline was not effective in reducing RMSD with the operational test of length 20, nor with any of the simulated tests. There was, however, consistent improvement resulting from the use of the smoothing splines as measured by RMSD. This divergence of measures of effectiveness suggests the need for further study before unequivocal recommendations may be made.

Presmoothers other than the negative hypergeometric are either ineffective, inconsistent in their effects, or have associated with them disadvantages such as greatly increased ASD. Divgi (1983) likewise found merit in the use of the negative hypergeometric, although he also found that the three-and four-parameter beta binomial distributions were more effective than the negative hypergeometric. (The negative hypergeometric is a two-parameter beta binomial.)

The lack of effectiveness of the other presmoothers may say less about the presmoothers than it does about the robustness of equipercentile equating. The various cumulative frequency counts used in equating may be degraded by all of the smoothers except the negative hypergeometric.

The cubic smoothing spline has a number of intuitively appealing characteristics: it can follow a curve of any shape, it can pass as close to the fit points as appropriate, and it is theoretically neutral in the sense that its use does not depend on the applicability or appropriateness of any statistical theory of testing. Its effectiveness, which is also reported by Kolen (1983), is thus not surprising. Although the improvements due to the splines were modest, the fact that there is no concomitant increase in ASD makes their use particularly attractive. The cubic smoothing splines perform, in effect, exactly what hand smoothing was to do: It passes a theoretically neutral curve among the points. Its effectiveness may derive from its mimicking of the original objective of postsmoothing.

Limitations

The present study is limited in several respects, all of which may tend to reduce its generalizability to other applications.

First, only five tests were used: two operational and three simulated. Generalizations to other tests may be inadvisable if the tests do not statistically resemble those used for this study.

Second, the tests used, especially the simulated tests, may be more similar to each other than are most operationally equated tests. Generalization to less similar tests is of questionable appropriateness.

Third, all equated pairs were pairs of tests of the same length, a condition not always found operationally.

Finally, within the current methodology it has not been possible to investigate a question of potential importance. One of the particularly significant advantages of equipercentile test equating is that when tests equated by the equipercentile method are used interchangeably to select only those who score at or above a certain percentile, then there is no expected advantage to any examinee in taking any particular form of the test in place of any other form. It is not clear that presmoothed equipercentile equatings retain that property. In applications where such percentile invariance is an essential consideration, the use of presmoothing should await further research.

Recommendations

Among the presmoothing methods, the negative hypergeometric and, by extension, other smoothers of the same beta binomial family, deserve consideration for operational use. If any of the presmoothers studied here is to be adopted, then the negative hypergeometric would be the most appropriate. It has the effect of reducing RMSD by about ten percent, a benefit which could also be achieved by increasing sample size by about 20 percent.

Among the postsmoothers, gains were not as evident with linear, quadratic, and cubic regression smoothing as had been anticipated. In those cases where an a priori decision has been made that the smoothing shall be linear, the use of orthogonal regression should be favored over the use of standard regression. Where the shape of the regression fitting is not determined in advance, then the use of cubic splines appears appropriate. These two postsmoothing methods, orthogonal regression and cubic splines, are appropriate for operational use with tests similar to those studied here, and may be useful with other tests if further research confirms their usefulness.

It is further suggested that future investigations consider not only AAD, RMSD, and ASD, but also look at the worst cases to determine whether, as Divgi (1983) found, there are some equatings in which the action of a generally helpful smoother results in less accurate results than does either no smoother or some other smoother. Divgi's results suggested that the four-parameter beta binomial was effective in most equatings, but that in a small proportion of cases its use was not appropriate because it increased deviations markedly.

Analytic derivations of the standard errors associated with equating distributions presmoothed by means of the negative hypergeometric or related smoothers would contribute to an understanding of the expected functioning of that smoother.

Finally, the figures and tables show that benefits are generally achieved at the cost of increases in ASD. The use of resampling techniques, such as jackknifing, known to reduce ASD, might be applied with the promising smoothers to determine if the two together would reduce RMSD and ASD. Such combining of a smoothing technique with a resampling might bring the benefits of both to the equating process.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 508-600.

Cureton, E. E., & Tukey, J. W. (1951). Smoothing frequency distributions, equating tests, and preparing norms. (Abstract of presented paper.) *American Psychologist*, *6*, 404.

Divgi, D. R. (1983). *Comparison of some methods for smoothing score distributions*. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic Press.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, *27*, 59-72.

Kolen, M. J. (1983). *Effectiveness of analytic smoothing in equipercentile equating*. (ACT Technical Bulletin Number 41), Iowa City, IA: American College Testing Program.

Lindsay, C. A., & Prichard, M. A. (1974). An analytical procedure for the equipercentile method of equating tests. *Journal of Educational Measurement*, *8*, 203-207.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1982). The standard error of equipercentile test equating. *Journal of Educational Statistics*, *7*, 165-174.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1983). *Comparison of IRT observed score and true score equatings*. (RR-83-26-ONR), Princeton, NJ: Educational Testing Service.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, *54*, 173-205.

Mosteller F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Ree, M. J. (1979). Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-385.

Ree, M. J. (1980). AVRAM: Adaptive vector and response automation method. Applied Psychological Measurement, 4, 277-278.

Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. (1982). Armed Services Vocational Aptitude Battery: Item and factor analysis of forms 8, 9, and 10 (AFHRL-TR-81-55, AO-A113 465). Brooks AFB, TX: Air Force Human Resources Laboratory.

Reinsch, C. H. (1967). Smoothing by spline functions. Numerische Mathematik, 10, 177-183.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

United States Military Entrance Processing Command (1984). Test manual for the Armed Services Vocational Aptitude Battery. North Chicago, IL: Document number DoD 1304.12AA, Author.

Velleman, P. F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. Journal of the American Statistical Association, 75, 609-615.

Velleman, P. F., & Hoaglin, D. C. (1981). Applications, basics, and computing of exploratory data analysis. Belmont, CA: Duxbury Press division of Wadsworth.

Appendix A

Technical Description of Simulated Tests

The tests used in the simulated equatings are described in detail in the accompanying tables. Item statistics, distribution statistics, and item parameters are all given. Test 15.1 was the experimental test of length 15, and test 15.2 was the reference test. The tests of lengths 30 and 50 were numbered similarly. The item parameters (a, b, and c) were used to develop the tests and to derive the expected observed score distributions. All of the other statistics for the items and for the test as a whole were developed from a single sample of 2,000 simulees. Table A-1 presents test summary statistics. Tables A-2 through A-7 present item statistics.

Table A - 1. Test Statistics for the Simulated Tests

Test	Mean	Standard Deviations	Skew	Kurtosis	KR-20	KR-21
15.1	10.25	3.28	-0.44	-0.67	.80	.75
15.2	10.55	3.25	-0.48	-0.67	.80	.75
30.1	18.54	6.63	-0.14	-0.97	.89	.87
30.2	17.93	6.57	0.00	-0.95	.88	.86
50.1	36.33	10.36	-0.57	-0.66	.93	.93
50.2	35.17	9.61	-0.48	-0.67	.91	.91

Table A - 2. Item Statistics for Tests 15.1

ITEM	A	B	C	P	RP	RB
1	1.067	-2.146	0.237	0.950	0.328	0.690
2	1.222	-1.290	0.240	0.883	0.481	0.788
3	1.297	-1.120	0.230	0.868	0.492	0.778
4	1.420	-0.793	0.273	0.816	0.526	0.765
5	1.487	-0.757	0.247	0.804	0.541	0.776
6	1.330	-0.591	0.240	0.766	0.564	0.779
7	1.900	-0.395	0.250	0.722	0.624	0.833
8	1.417	-0.350	0.260	0.736	0.561	0.756
9	1.439	-0.265	0.261	0.717	0.586	0.780
10	1.420	-0.094	0.241	0.637	0.583	0.747
11	1.371	0.007	0.269	0.656	0.543	0.701
12	1.590	0.677	0.226	0.449	0.499	0.628
13	1.150	0.682	0.245	0.489	0.489	0.613
14	1.530	0.778	0.240	0.429	0.478	0.603
15	1.430	1.393	0.235	0.326	0.356	0.463

KEY: A = a parameter; B = b parameter; C = c parameter; P = proportion of simulees answering correctly; RP = point biserial correlation; RB = biserial item-test correlation.

Table A -3. Item Statistics for Tests 15.2

ITEM	A	B	C	P	RP	RB
1	1.557	-1.940	0.248	0.956	0.339	0.746
2	1.413	-1.790	0.234	0.944	0.384	0.780
3	1.240	-1.226	0.237	0.878	0.468	0.756
4	1.080	-0.866	0.237	0.804	0.502	0.720
5	1.410	-0.614	0.242	0.768	0.569	0.787
6	1.370	-0.488	0.248	0.752	0.577	0.788
7	1.220	-0.460	0.256	0.729	0.546	0.733
8	1.096	-0.412	0.251	0.735	0.526	0.709
9	1.350	-0.340	0.248	0.728	0.584	0.783
10	1.468	-0.299	0.215	0.680	0.620	0.809
11	1.405	-0.140	0.244	0.675	0.561	0.730
12	1.670	0.146	0.246	0.584	0.581	0.734
13	1.494	0.482	0.227	0.508	0.538	0.674
14	1.263	0.865	0.262	0.440	0.454	0.571
15	1.146	1.261	0.247	0.368	0.375	0.480

KEY: A = a parameter; B = b parameter; C = c parameter; P = proportion of simulees answering correctly; RP = point biserial correlation; RB = biserial item-test correlation.

Table A -4. Item Statistics for Tests 30.1

ITEM	A	B	C	P	RP	RB
1	1.706	-1.800	0.242	0.944	0.362	0.737
2	1.051	-1.177	0.242	0.850	0.441	0.675
3	1.362	-1.088	0.229	0.852	0.519	0.798
4	1.639	-1.011	0.231	0.846	0.509	0.774
5	1.447	-0.926	0.217	0.820	0.537	0.786
6	1.771	-0.838	0.223	0.827	0.555	0.821
7	1.411	-0.806	0.240	0.806	0.521	0.749
8	1.262	-0.758	0.217	0.792	0.515	0.729
9	1.291	-0.683	0.244	0.796	0.527	0.749
10	1.379	-0.520	0.222	0.739	0.554	0.749
11	1.108	-0.380	0.227	0.715	0.519	0.690
12	1.443	-0.241	0.227	0.692	0.582	0.764
13	1.685	-0.064	0.224	0.640	0.593	0.761
14	1.403	-0.023	0.231	0.624	0.571	0.729
15	1.673	0.003	0.235	0.617	0.593	0.755
16	1.403	0.077	0.225	0.596	0.544	0.689
17	1.283	0.217	0.227	0.559	0.489	0.615
18	1.377	0.323	0.221	0.542	0.518	0.651
19	1.656	0.420	0.219	0.503	0.540	0.676
20	1.641	0.512	0.216	0.472	0.506	0.634
21	1.603	0.563	0.225	0.478	0.511	0.641
22	1.287	0.569	0.222	0.488	0.485	0.608
23	1.226	0.579	0.226	0.486	0.453	0.567
24	1.242	0.611	0.249	0.499	0.439	0.550
25	1.313	0.624	0.229	0.467	0.473	0.594
26	0.930	0.880	0.235	0.437	0.421	0.530
27	1.266	0.895	0.221	0.407	0.384	0.486
28	1.159	0.994	0.224	0.406	0.388	0.491
29	1.578	1.251	0.215	0.323	0.397	0.517
30	1.629	1.490	0.243	0.315	0.290	0.379

KEY: A = a parameter; B = b parameter; C = c parameter; P = proportion of simulees answering correctly; RP = point biserial correlation; RB = biserial item-test correlation.

Table A -5. Item Statistics for Tests 30.2

ITEM	A	B	C	P	RP	RB
1	1.266	-1.498	0.239	0.907	0.379	0.661
2	1.717	-1.424	0.232	0.924	0.364	0.676
3	1.287	-1.110	0.236	0.856	0.464	0.717
4	1.010	-0.731	0.236	0.776	0.453	0.632
5	1.373	-0.675	0.202	0.775	0.533	0.741
6	1.336	-0.613	0.243	0.771	0.505	0.701
7	0.977	-0.605	0.243	0.755	0.458	0.627
8	1.450	-0.497	0.224	0.747	0.553	0.752
9	1.310	-0.264	0.235	0.673	0.518	0.674
10	1.211	-0.169	0.221	0.666	0.549	0.711
11	1.349	-0.150	0.226	0.660	0.542	0.700
12	1.100	-0.131	0.230	0.661	0.523	0.676
13	1.509	-0.086	0.217	0.660	0.577	0.746
14	1.361	-0.065	0.229	0.625	0.553	0.706
15	1.240	-0.032	0.242	0.613	0.518	0.659
16	1.337	0.020	0.212	0.622	0.563	0.719
17	1.671	0.082	0.244	0.613	0.566	0.720
18	1.663	0.198	0.236	0.558	0.580	0.729
19	1.573	0.227	0.226	0.579	0.596	0.752
20	1.606	0.357	0.252	0.540	0.533	0.669
21	1.418	0.364	0.240	0.548	0.534	0.671
22	1.591	0.678	0.223	0.458	0.486	0.611
23	1.601	0.898	0.223	0.410	0.480	0.607
24	1.349	1.001	0.211	0.385	0.412	0.524
25	1.557	1.061	0.241	0.399	0.416	0.528
26	1.547	1.114	0.255	0.397	0.374	0.475
27	1.124	1.287	0.238	0.378	0.377	0.481
28	1.307	1.304	0.234	0.363	0.309	0.396
29	1.697	1.548	0.230	0.305	0.315	0.414
30	1.900	1.600	0.235	0.304	0.292	0.384

KEY: A = a parameter; B = b parameter; C = c parameter; P = proportion of simulees answering correctly; RP = point biserial correlation; RB = biserial item-test correlation.

Table A - 6. **Item Statistics for Tests 50.1**

ITEM	A	B	C	P	RP	RB
1	1.330	-1.620	0.229	0.921	0.429	0.788
2	1.538	-1.540	0.254	0.923	0.451	0.830
3	1.433	-1.410	0.229	0.897	0.490	0.832
4	0.770	-1.311	0.244	0.840	0.404	0.609
5	1.005	-1.265	0.250	0.856	0.431	0.668
6	1.096	-1.237	0.249	0.857	0.481	0.746
7	0.900	-1.225	0.253	0.840	0.440	0.663
8	0.951	-1.171	0.240	0.839	0.441	0.663
9	1.321	-1.158	0.248	0.877	0.483	0.780
10	1.080	-1.123	0.241	0.840	0.433	0.651
11	1.101	-1.086	0.255	0.842	0.464	0.701
12	0.828	-1.012	0.239	0.799	0.439	0.627
13	1.060	-0.984	0.249	0.818	0.496	0.724
14	1.222	-0.957	0.225	0.820	0.529	0.775
15	1.200	-0.950	0.228	0.816	0.510	0.742
16	0.822	-0.882	0.251	0.792	0.415	0.588
17	1.110	-0.861	0.240	0.792	0.512	0.726
18	1.171	-0.859	0.244	0.814	0.494	0.718
19	1.193	-0.786	0.226	0.785	0.513	0.721
20	1.383	-0.748	0.261	0.795	0.528	0.750
21	1.067	-0.739	0.247	0.772	0.488	0.678
22	1.014	-0.694	0.234	0.746	0.534	0.726
23	1.026	-0.559	0.240	0.734	0.510	0.687
24	1.220	-0.515	0.245	0.741	0.545	0.737
25	1.267	-0.487	0.248	0.726	0.513	0.686
26	0.979	-0.464	0.246	0.725	0.469	0.627
27	1.273	-0.434	0.263	0.731	0.509	0.683
28	1.156	-0.432	0.240	0.710	0.528	0.700
29	1.280	-0.429	0.240	0.722	0.569	0.760
30	1.401	-0.427	0.234	0.721	0.557	0.743

Table A - 6. (continued)

31	1.243	-0.426	0.246	0.721	0.521	0.695
32	1.098	-0.425	0.241	0.720	0.495	0.660
33	1.034	-0.279	0.240	0.687	0.475	0.621
34	0.624	-0.219	0.252	0.684	0.332	0.434
35	1.465	-0.216	0.235	0.668	0.575	0.746
36	1.133	-0.170	0.245	0.657	0.545	0.703
37	1.074	-0.167	0.232	0.661	0.473	0.612
38	1.394	-0.161	0.241	0.658	0.560	0.723
39	0.829	-0.157	0.250	0.655	0.414	0.533
40	1.366	-0.106	0.238	0.636	0.555	0.712
41	1.222	-0.104	0.256	0.664	0.512	0.663
42	0.959	0.018	0.239	0.605	0.464	0.589
43	1.381	0.072	0.229	0.572	0.555	0.700
44	1.069	0.318	0.247	0.549	0.450	0.565
45	0.981	0.343	0.232	0.538	0.427	0.536
46	1.389	0.362	0.226	0.541	0.495	0.622
47	1.166	0.415	0.251	0.540	0.441	0.554
48	1.071	0.496	0.251	0.514	0.434	0.544
49	1.477	0.517	0.235	0.490	0.451	0.565
50	1.268	0.605	0.227	0.473	0.437	0.548

KEY: A = a parameter; B = b parameter; C = c parameter; P = proportion of simulees answering correctly; RP = point biserial correlation; RB = biserial item-test correlation.

Table A - 7. Item Statistics for Tests 50.2

ITEM	A	B	C	P	RP	RB
1	0.976	-1.593	0.236	0.886	0.397	0.653
2	0.941	-1.367	0.255	0.870	0.400	0.636
3	0.936	-1.268	0.233	0.863	0.433	0.678
4	1.224	-1.183	0.246	0.875	0.435	0.699
5	0.946	-1.177	0.234	0.843	0.414	0.627
6	0.890	-1.166	0.242	0.832	0.423	0.629
7	1.072	-1.141	0.263	0.843	0.485	0.734
8	0.966	-1.139	0.236	0.815	0.452	0.656
9	1.046	-1.127	0.230	0.837	0.485	0.728
10	1.142	-1.086	0.222	0.840	0.505	0.761
11	1.085	-1.019	0.241	0.843	0.476	0.722
12	0.867	-0.953	0.232	0.808	0.408	0.589
13	0.878	-0.942	0.240	0.799	0.440	0.628
14	0.816	-0.820	0.253	0.782	0.415	0.581
15	0.867	-0.799	0.229	0.785	0.462	0.650
16	0.882	-0.779	0.249	0.778	0.448	0.625
17	0.831	-0.670	0.227	0.743	0.476	0.645
18	0.830	-0.653	0.244	0.718	0.437	0.582
19	0.910	-0.567	0.239	0.737	0.482	0.650
20	0.940	-0.535	0.265	0.731	0.448	0.602
21	0.924	-0.508	0.247	0.737	0.491	0.662
22	1.129	-0.485	0.244	0.741	0.511	0.691
23	0.853	-0.418	0.238	0.696	0.430	0.565
24	1.116	-0.408	0.245	0.726	0.527	0.705
25	0.996	-0.395	0.243	0.710	0.447	0.592
26	0.951	-0.371	0.238	0.677	0.498	0.648
27	0.863	-0.333	0.222	0.685	0.429	0.561
28	0.783	-0.317	0.254	0.694	0.378	0.497
29	1.038	-0.201	0.230	0.669	0.472	0.613
30	0.913	-0.154	0.234	0.633	0.503	0.644

Table A - 7. (continued)

31	1.256	-0.077	0.235	0.623	0.516	0.659
32	0.921	-0.053	0.240	0.647	0.446	0.574
33	0.991	-0.041	0.238	0.627	0.492	0.629
34	0.532	-0.020	0.242	0.613	0.382	0.487
35	0.815	0.021	0.238	0.613	0.430	0.547
36	0.882	0.039	0.240	0.601	0.442	0.560
37	0.799	0.042	0.238	0.604	0.422	0.535
38	1.032	0.067	0.244	0.627	0.478	0.611
39	0.915	0.087	0.222	0.596	0.486	0.616
40	0.889	0.123	0.231	0.612	0.440	0.559
41	0.976	0.155	0.234	0.599	0.436	0.552
42	0.816	0.198	0.248	0.569	0.406	0.511
43	0.532	0.241	0.240	0.576	0.322	0.406
44	0.919	0.298	0.233	0.553	0.427	0.537
45	0.746	0.301	0.236	0.567	0.379	0.478
46	1.175	0.324	0.251	0.547	0.461	0.580
47	0.716	0.358	0.230	0.553	0.398	0.500
48	0.986	0.475	0.242	0.524	0.445	0.558
49	0.758	0.554	0.236	0.506	0.370	0.464
50	0.559	0.743	0.236	0.497	0.308	0.386

KEY: A = a parameter; B = b parameter; C = c parameter; P = proportion of simulees answering correctly; RP = point biserial correlation; RB = biserial item-test correlation.

Appendix B

Figures showing the effects of the smoothing procedures.

73

Figure B-1

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Presmoothed by 3-point moving medians

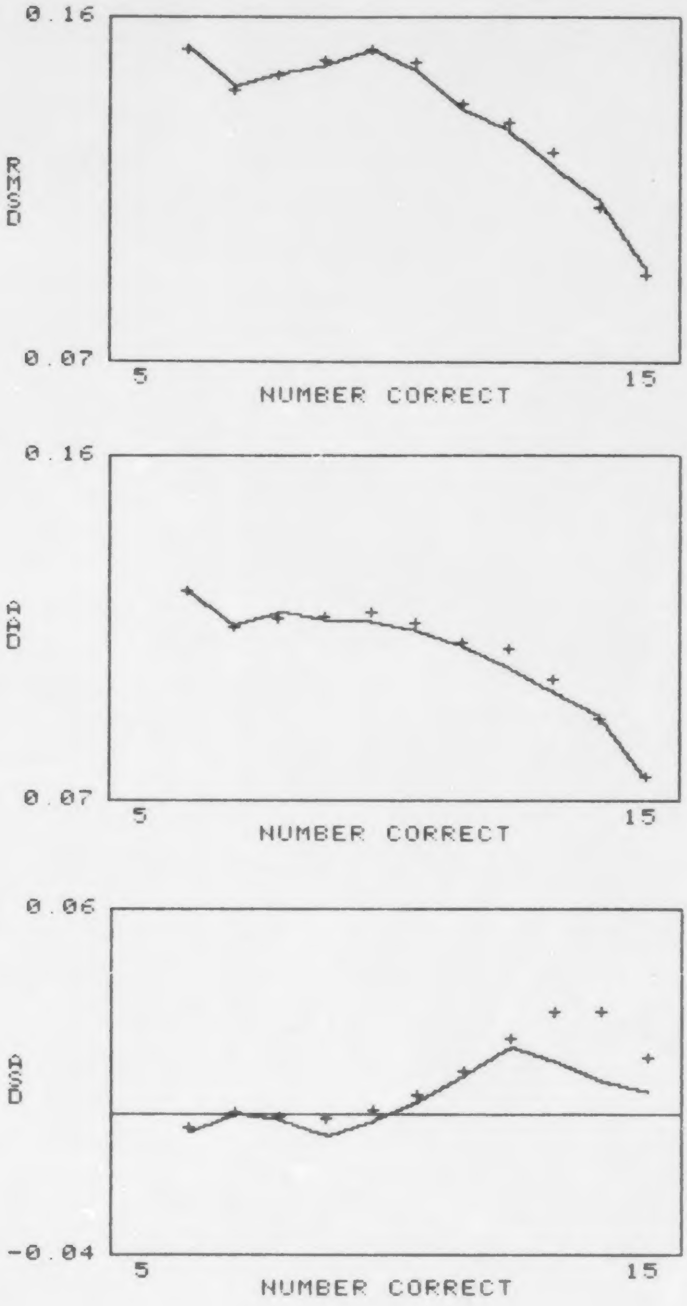


Figure B-2

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Presmoothed by 3-point moving medians

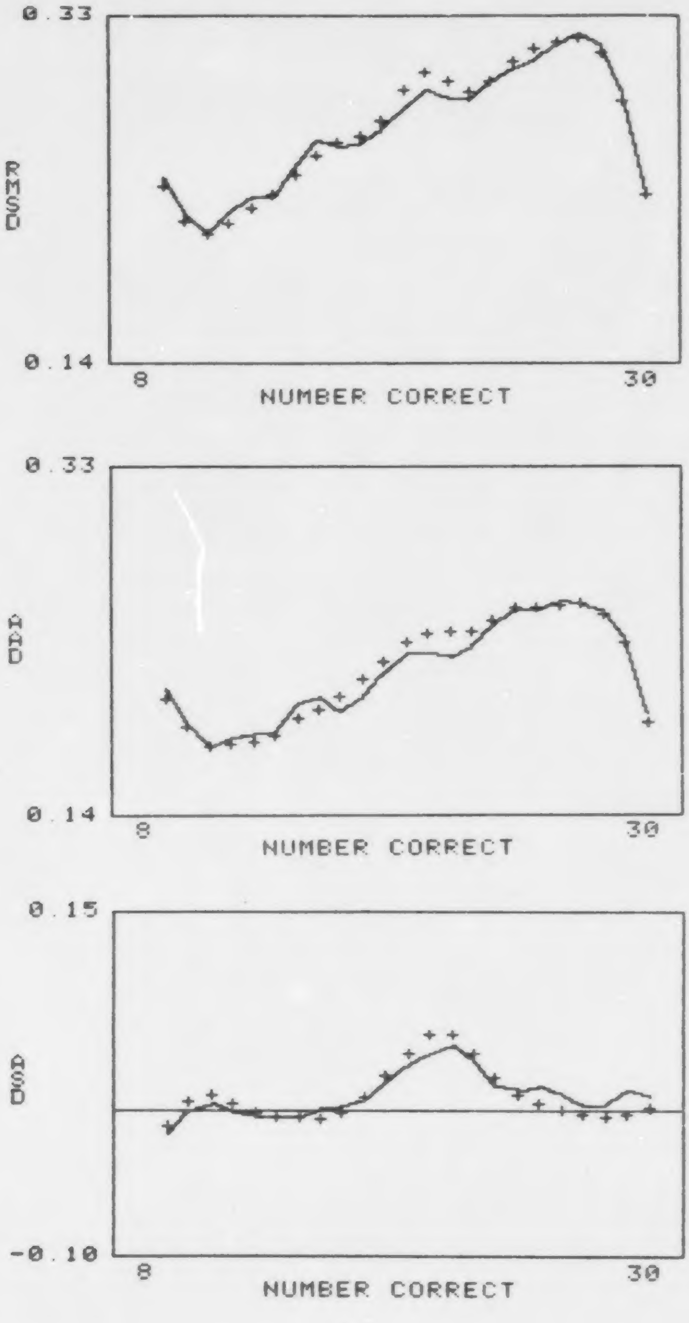


Figure B-3

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 50
Test Type: Simulated
Smoothing: Presmoothed by 3-point moving medians

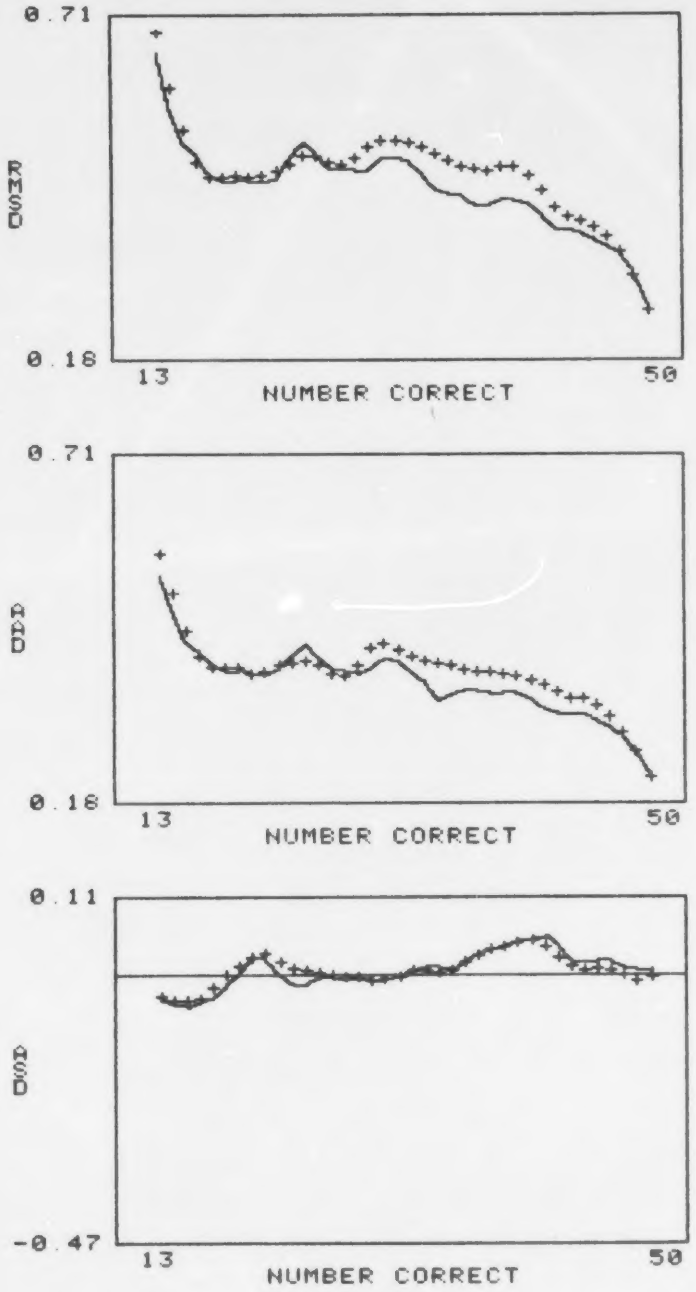


Figure B-4

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 20

Test Type: Operational

Smoothing: Presmoothed by 3-point moving medians

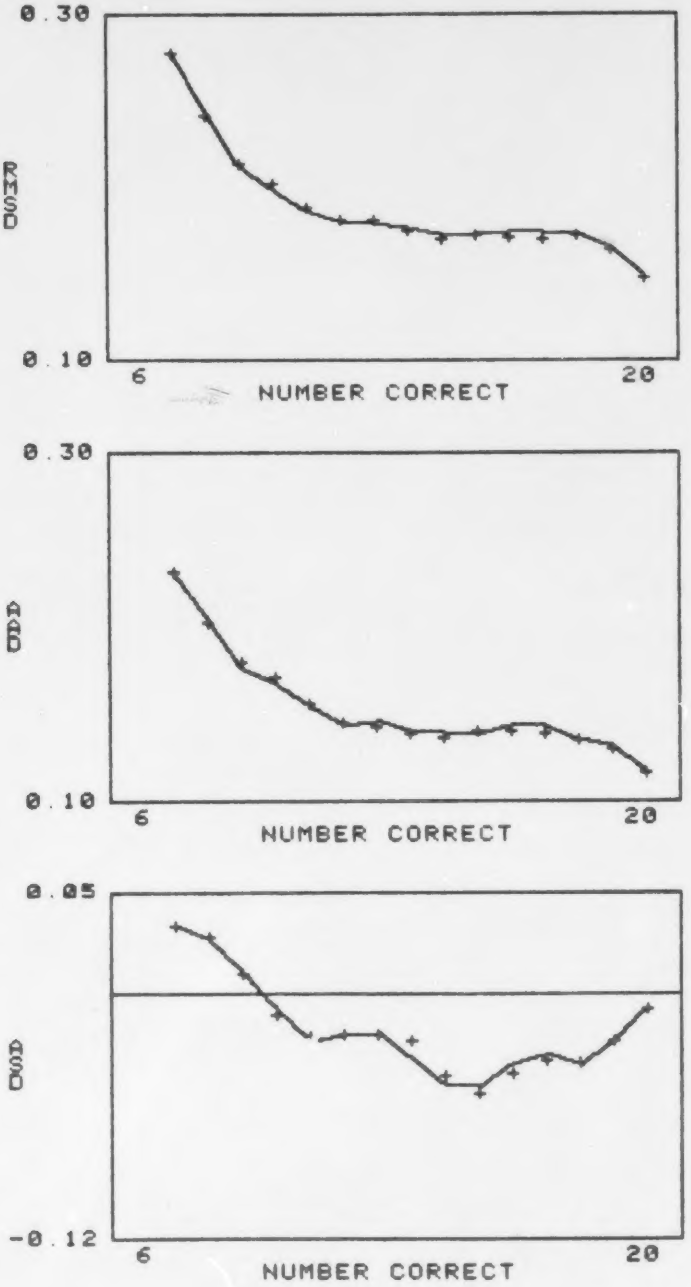


Figure B-5

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25
Test Type: Operational
Smoothing: Presmoothed by 3-point moving medians

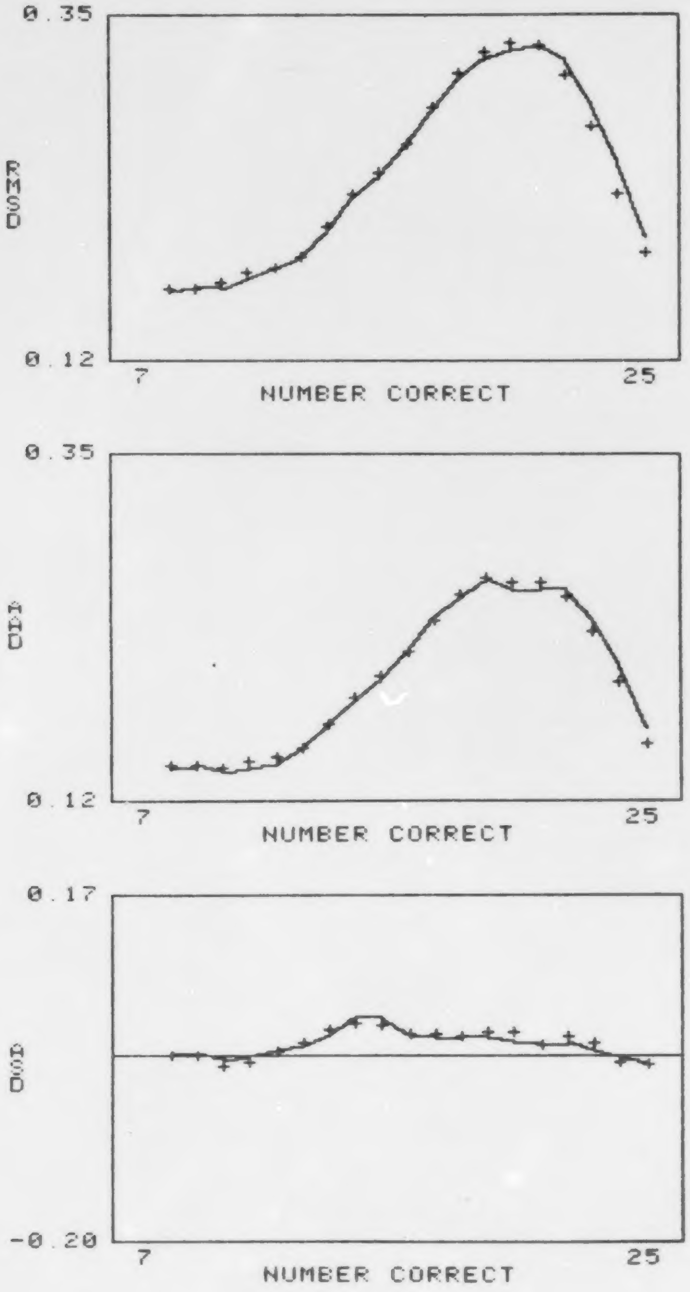


Figure B-6

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving medians

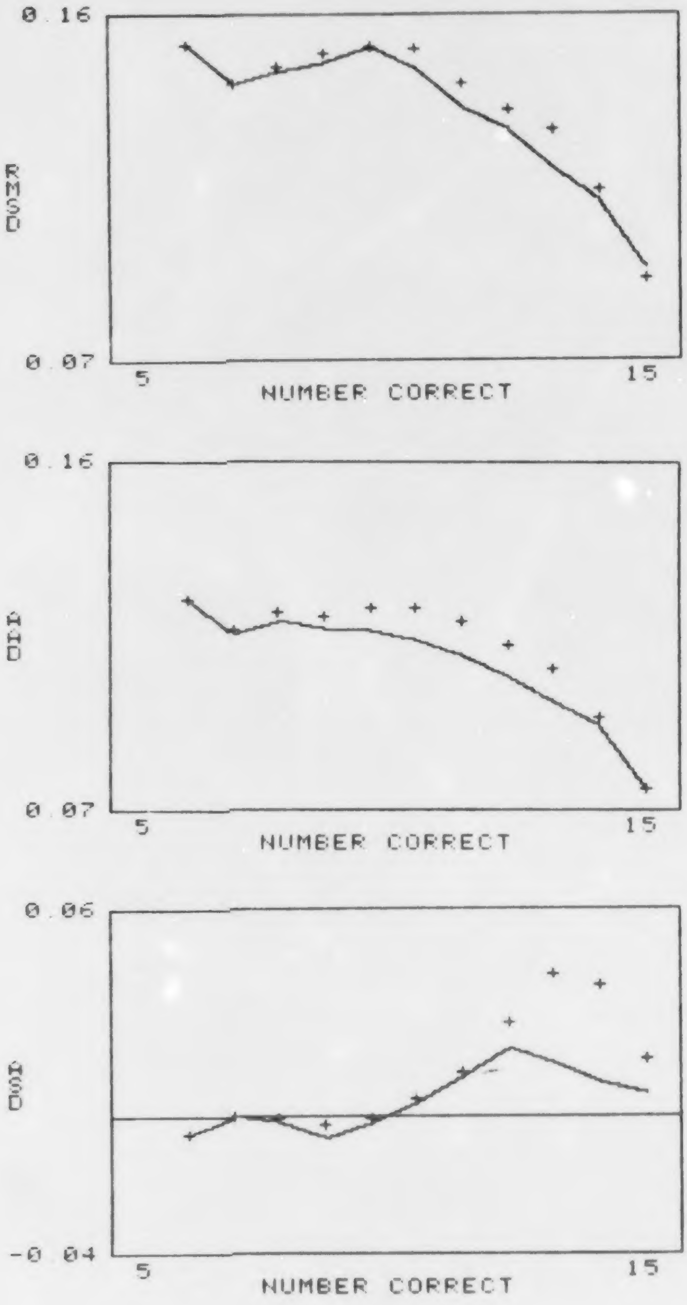


Figure B-7

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving medians

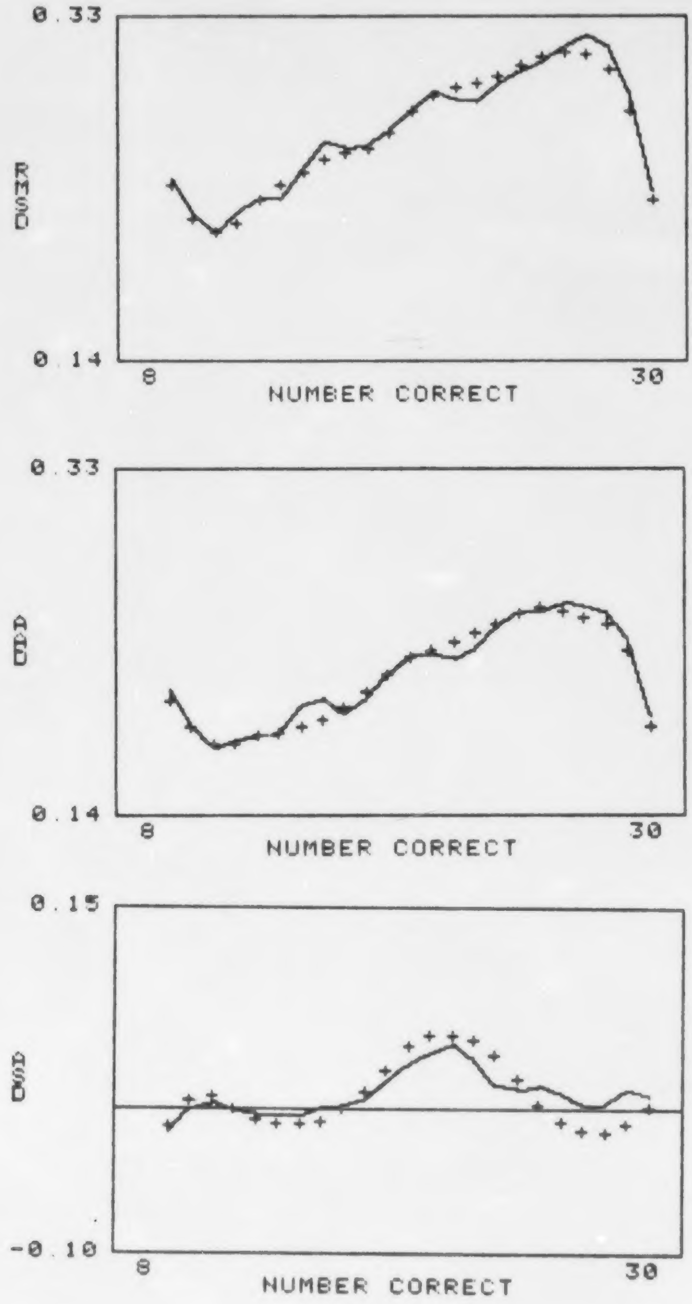


Figure B-8

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving medians

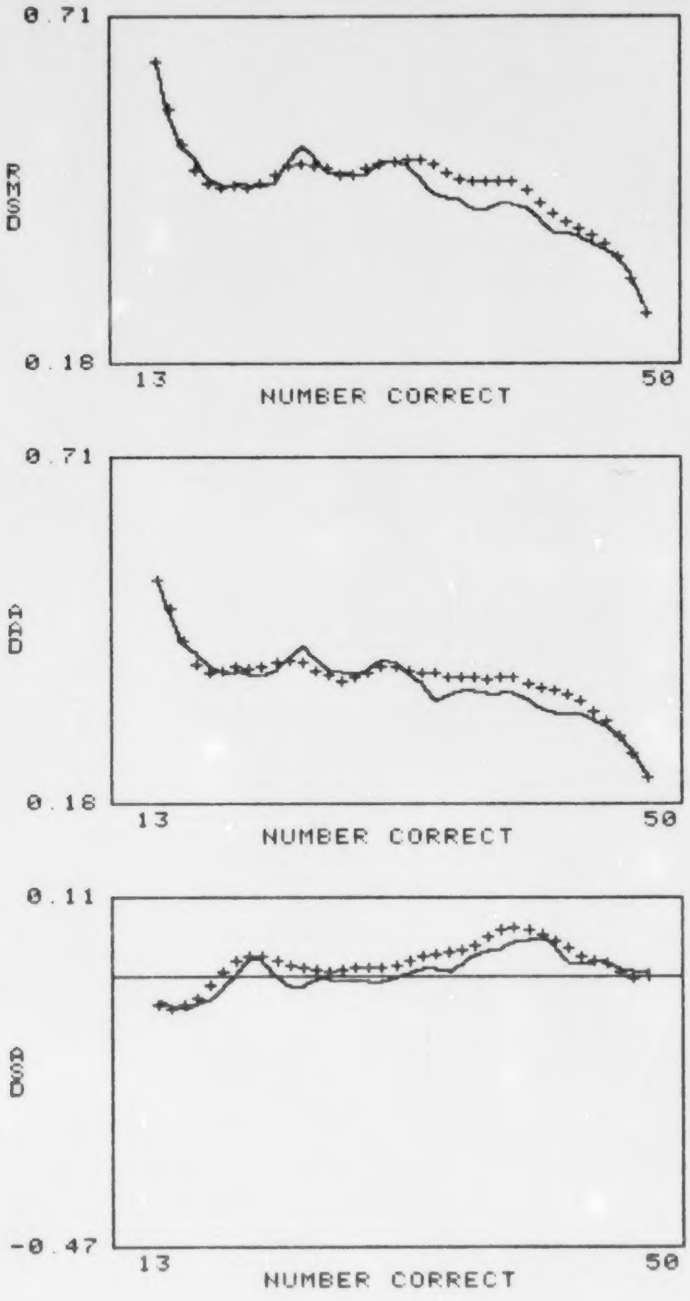


Figure B-9

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Presmoothed by 5-point moving medians

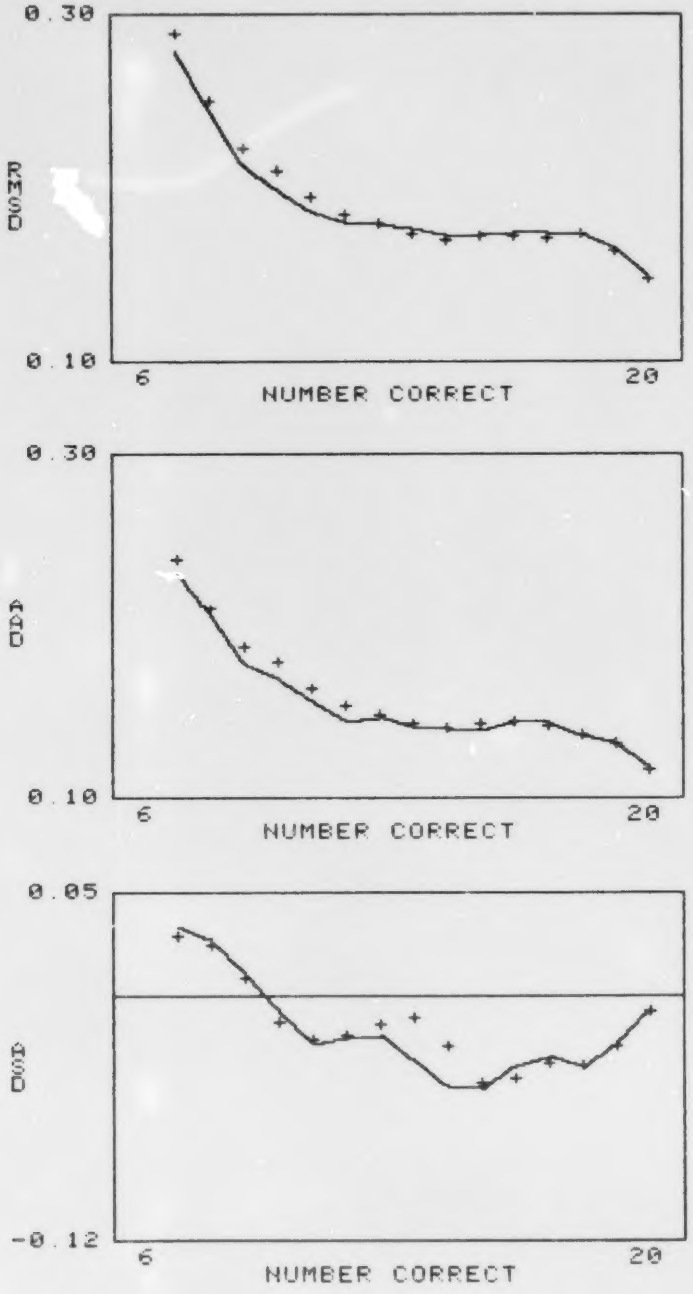


Figure B-10

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Presmoothed by 5-point moving medians

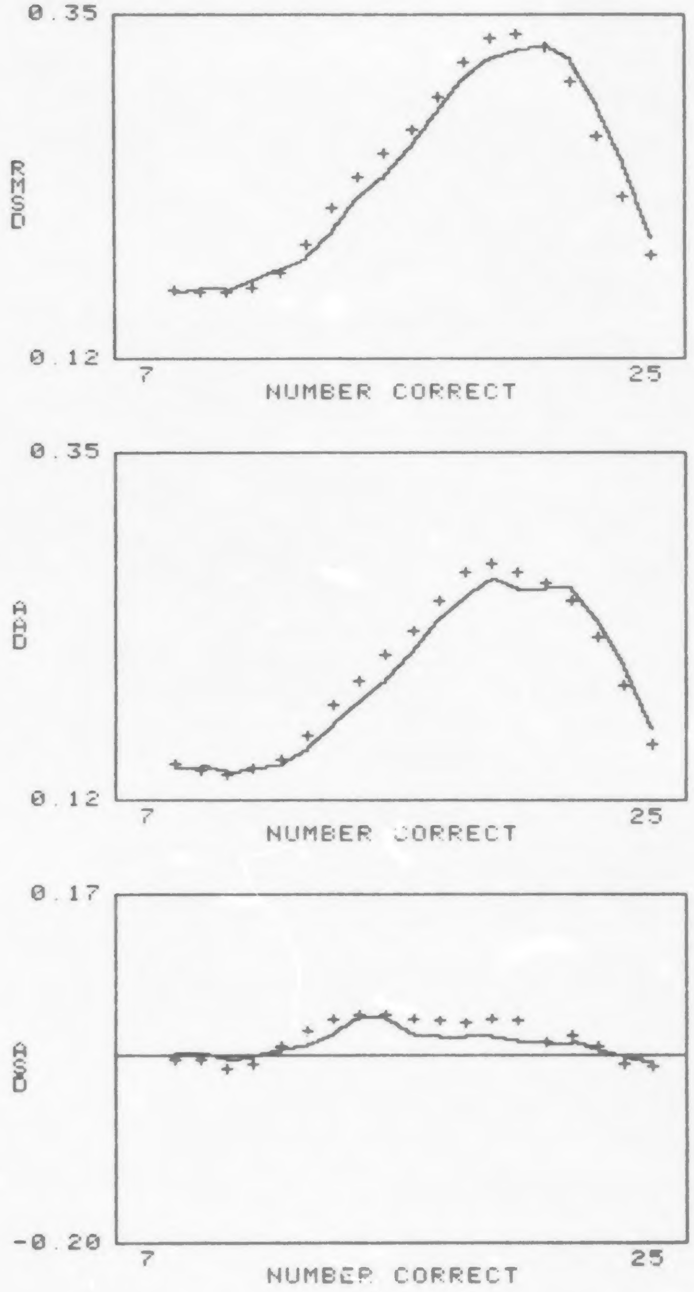


Figure B-11

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Presmoothed by 3-point moving weighted averages

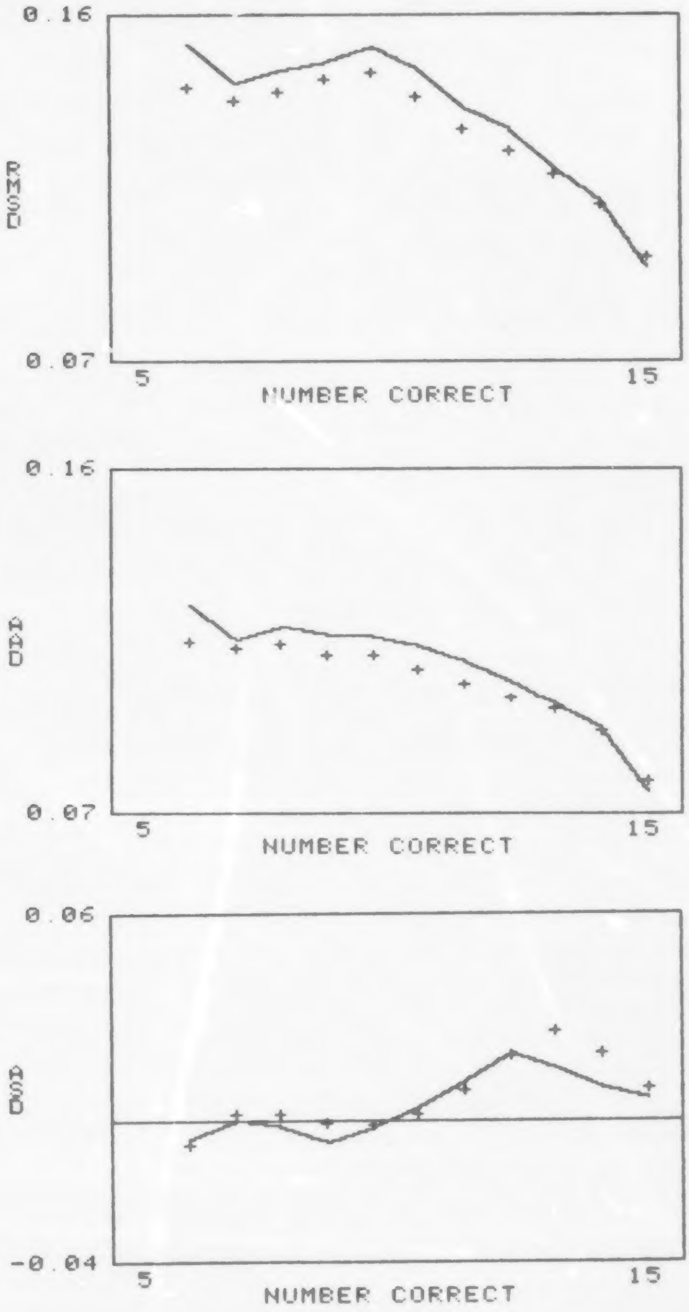


Figure B-12

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Presmoothed by 3-point moving weighted averages

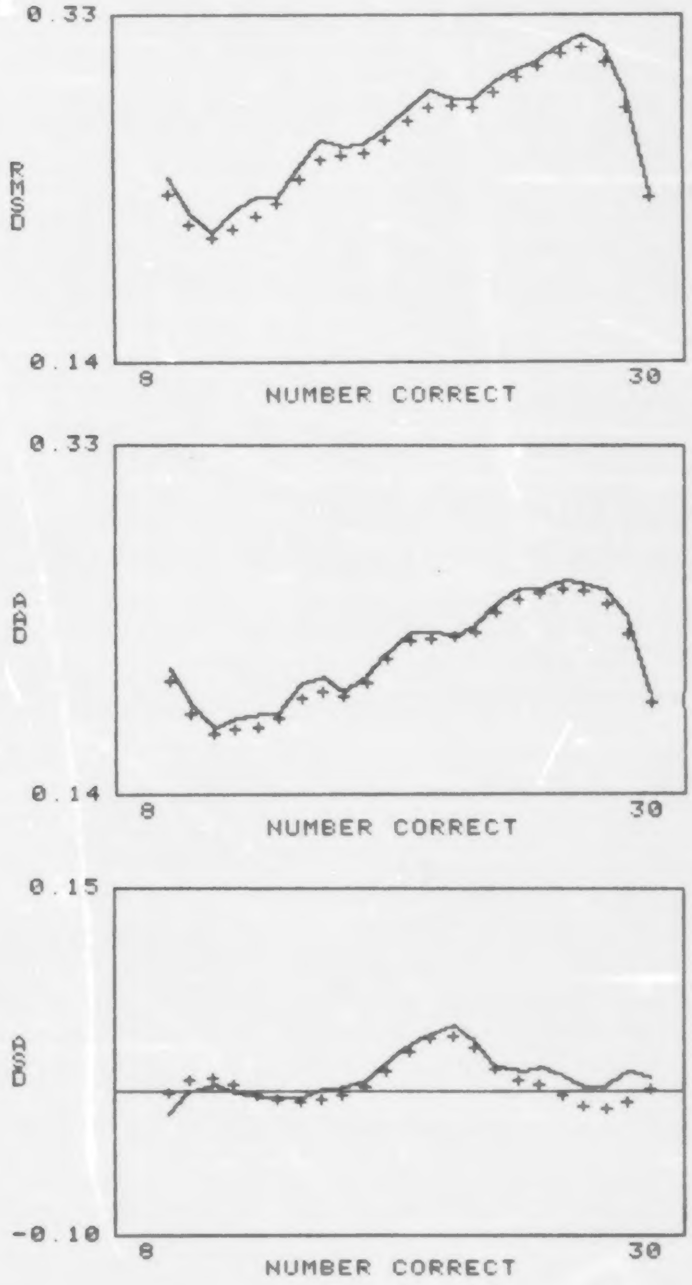


Figure B-13

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 50
Test Type: Simulated
Smoothing: Presmoothed by 3-point moving weighted averages

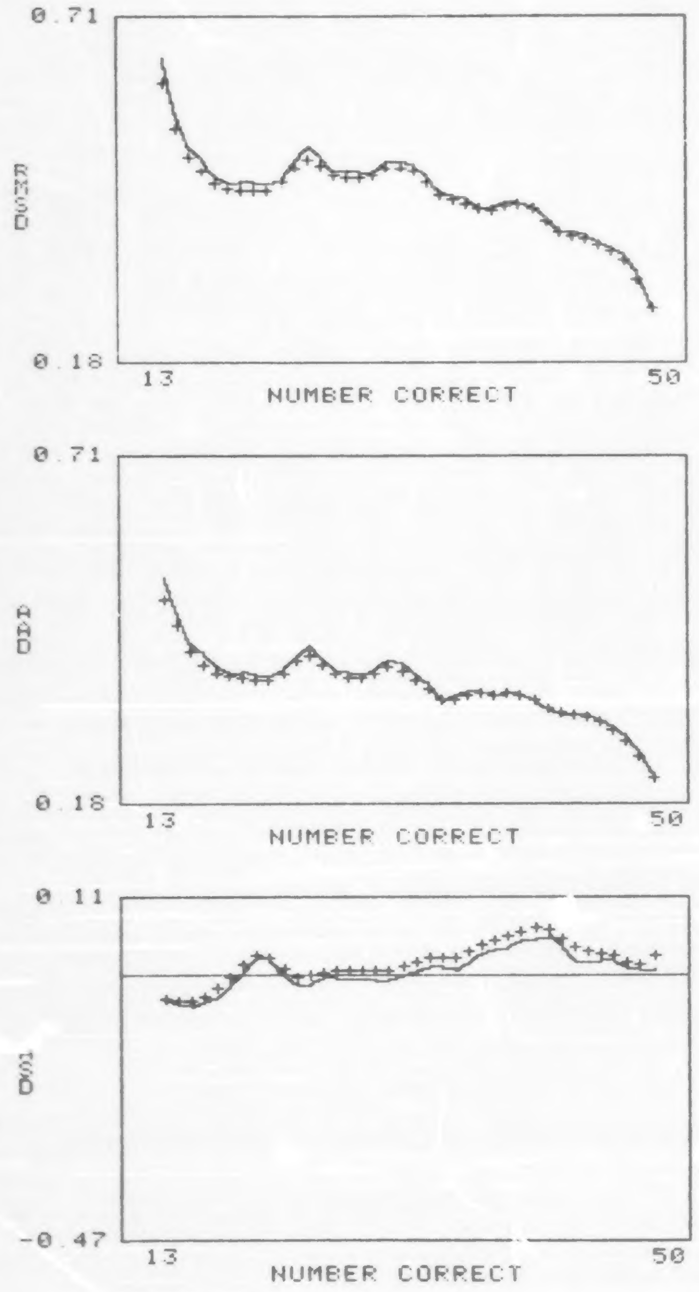


Figure B-14

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Presmoothed by 3-point moving weighted averages

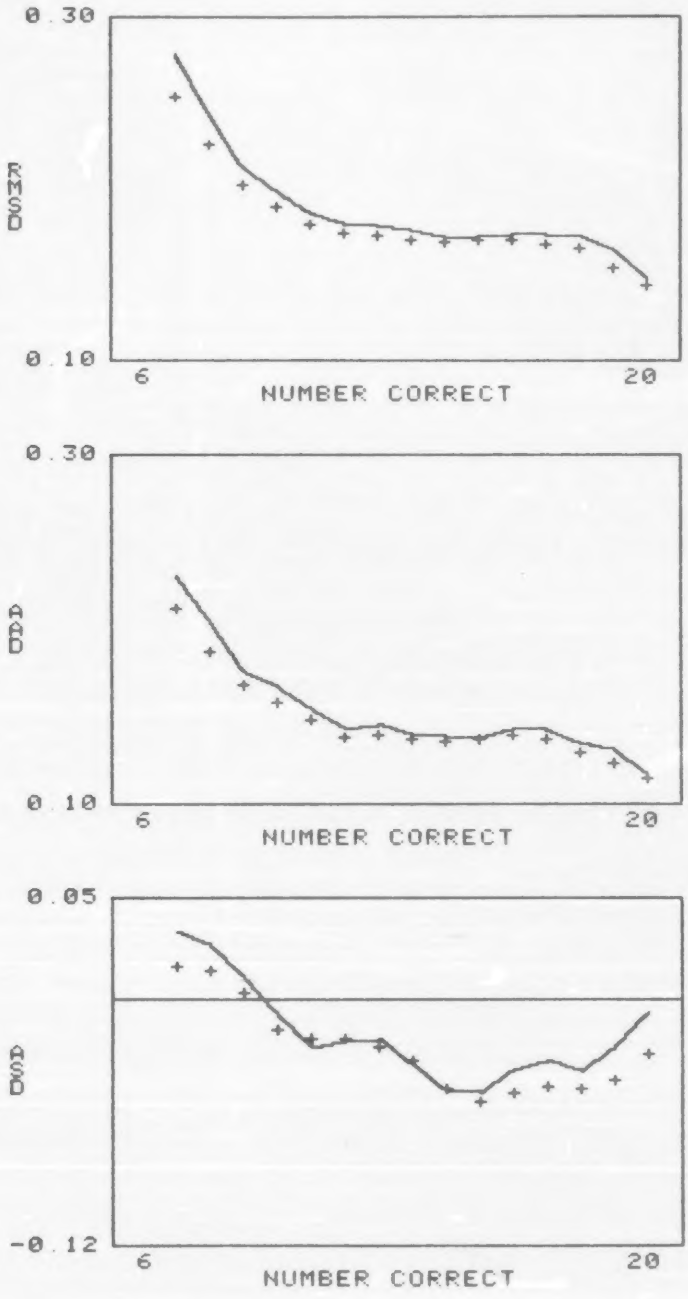


Figure B-15

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Presmoothed by 3-point moving weighted averages

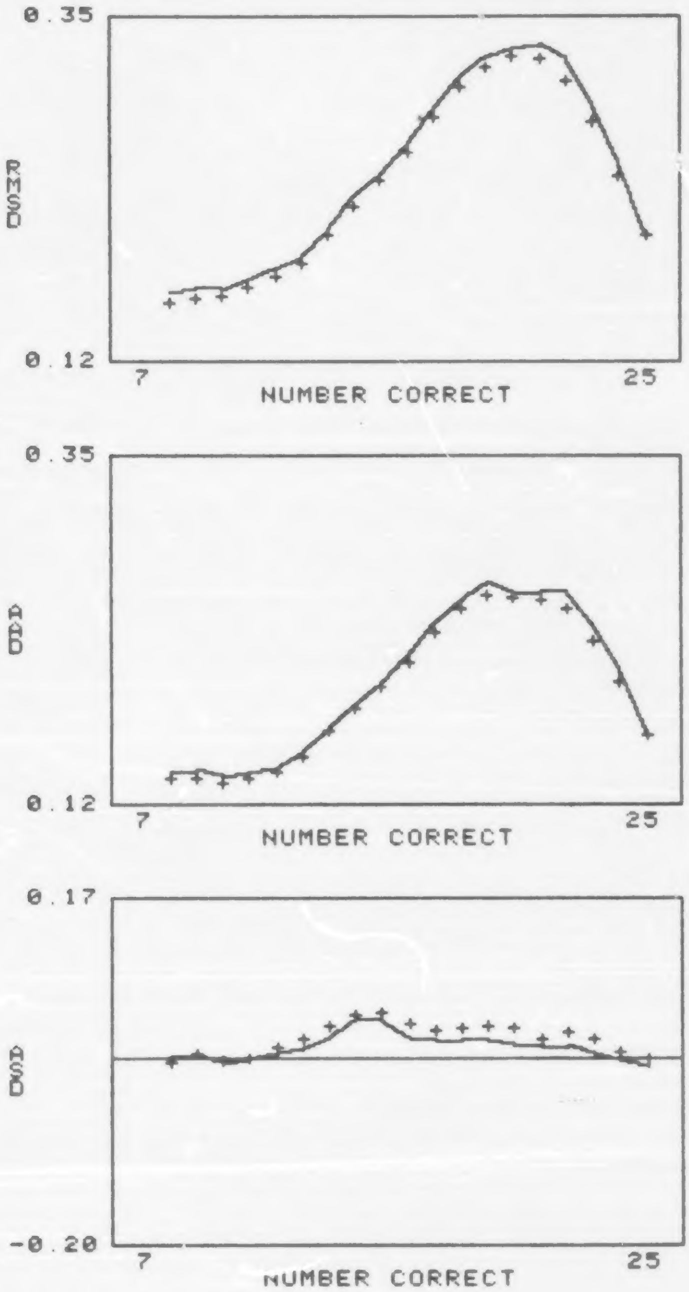


Figure B-16

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving weighted averages

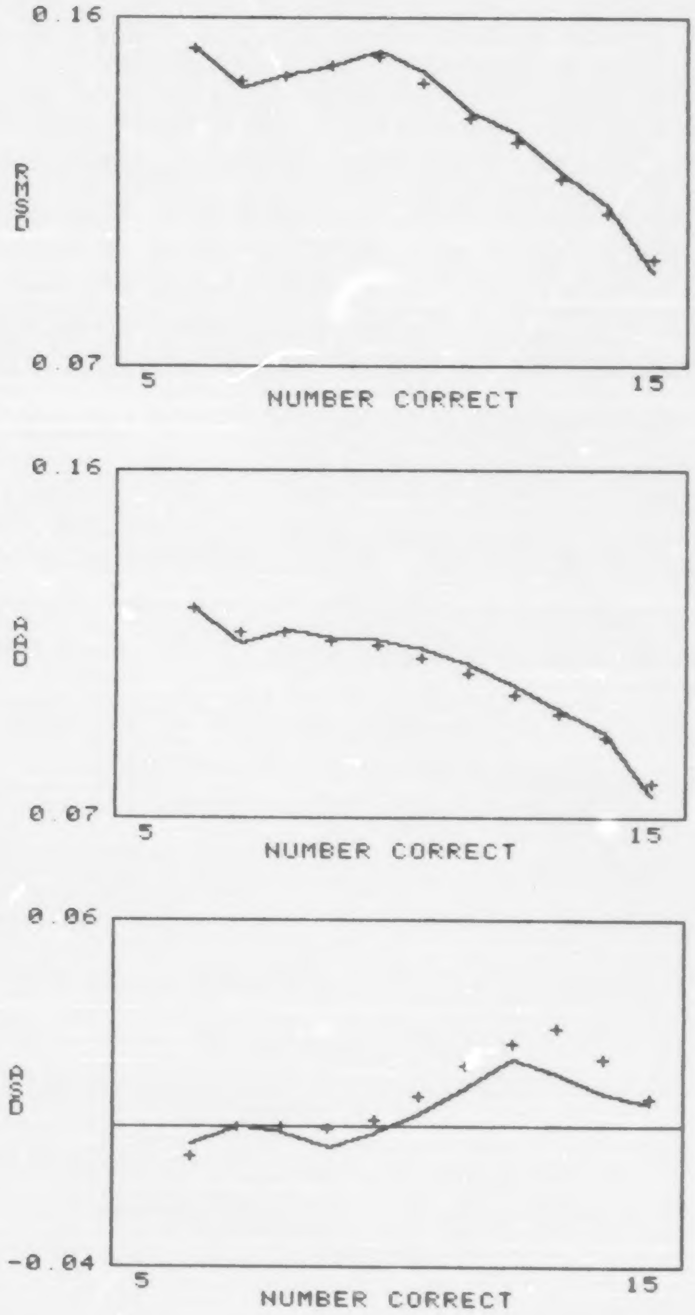


Figure B-17

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving weighted averages

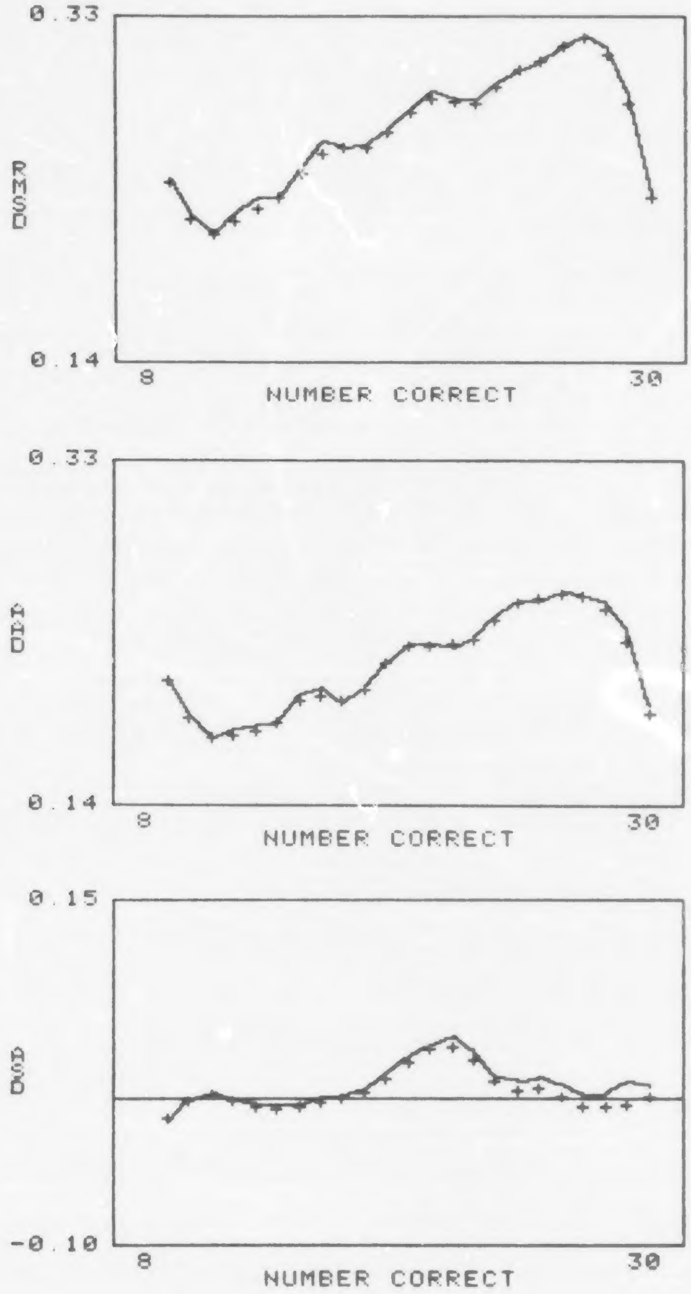


Figure B-18

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 50
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving weighted averages

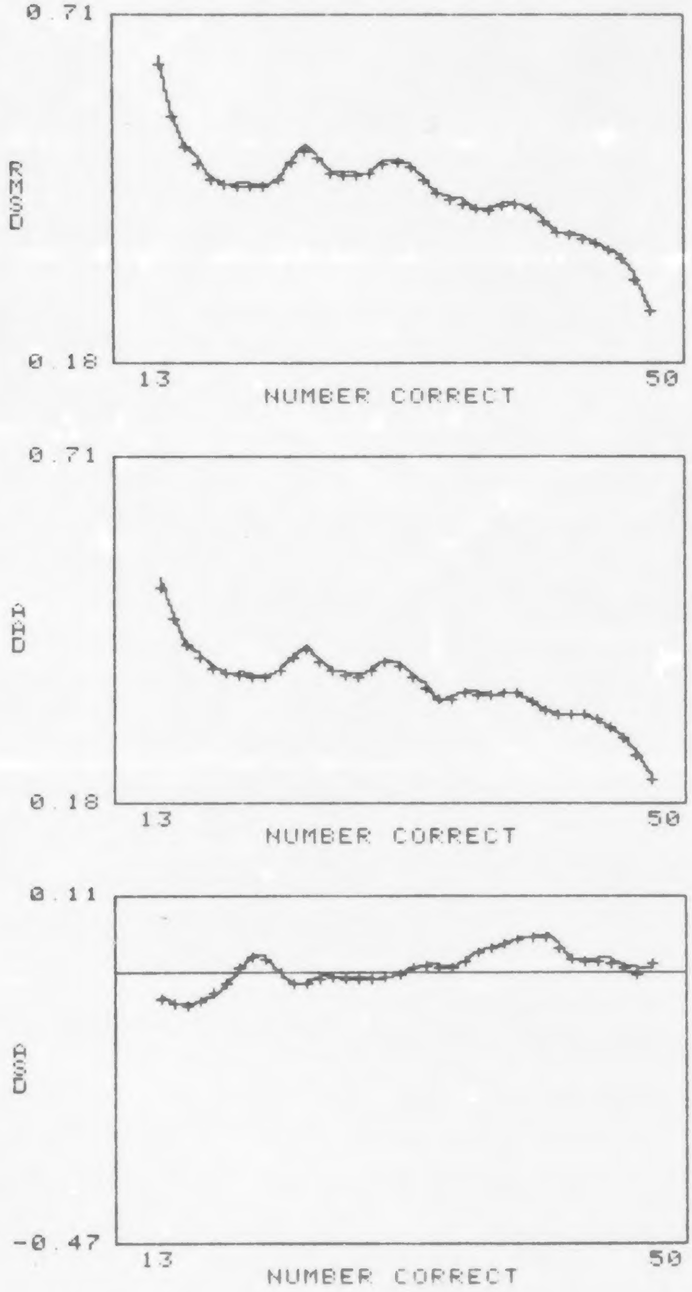


Figure B-19

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Presmoothed by 5-point moving weighted averages

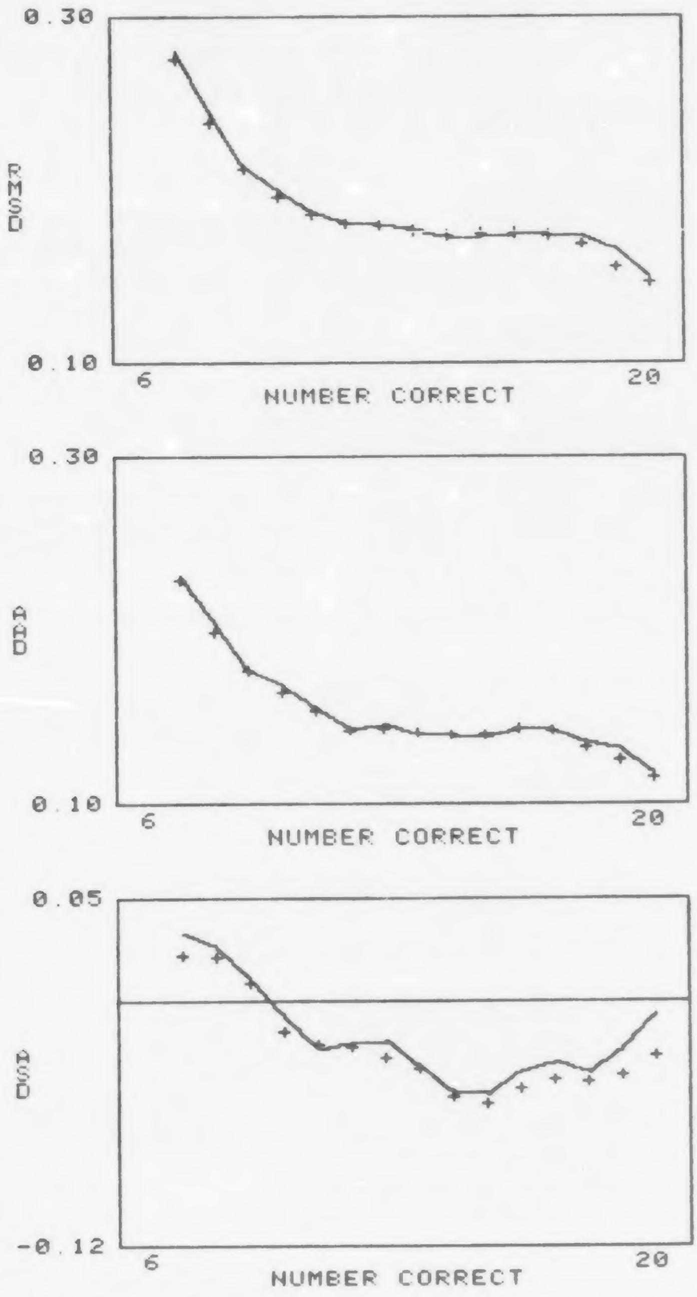


Figure B-20

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25
Test Type: Operational
Smoothing: Presmoothed by 5-point moving weighted averages

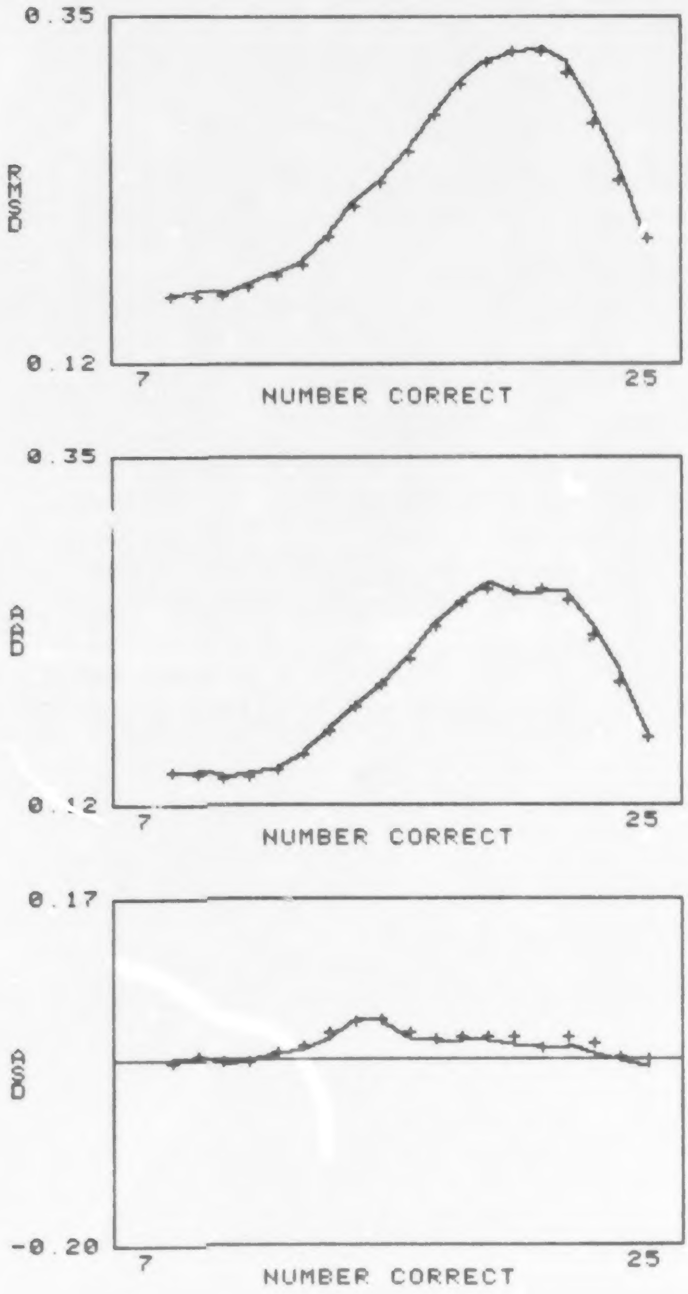


Figure B-21

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving weighted averages with root transformation

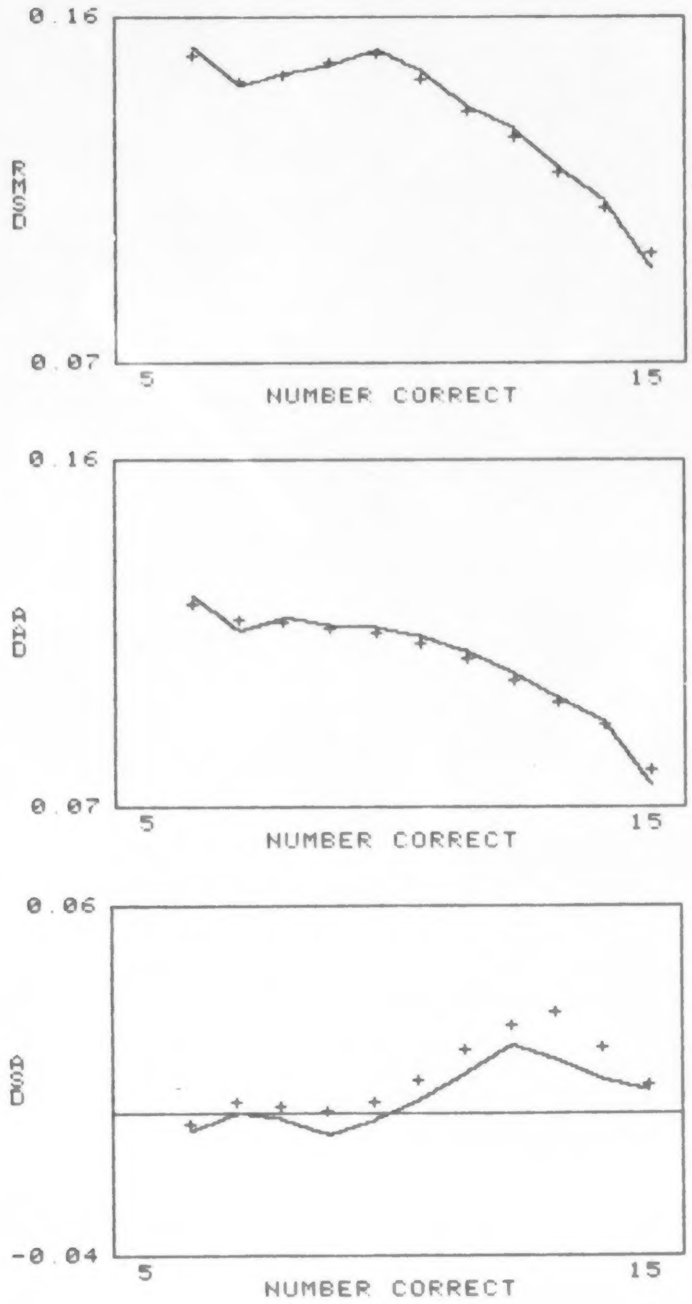


Figure B-22

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving weighted averages with root transformation

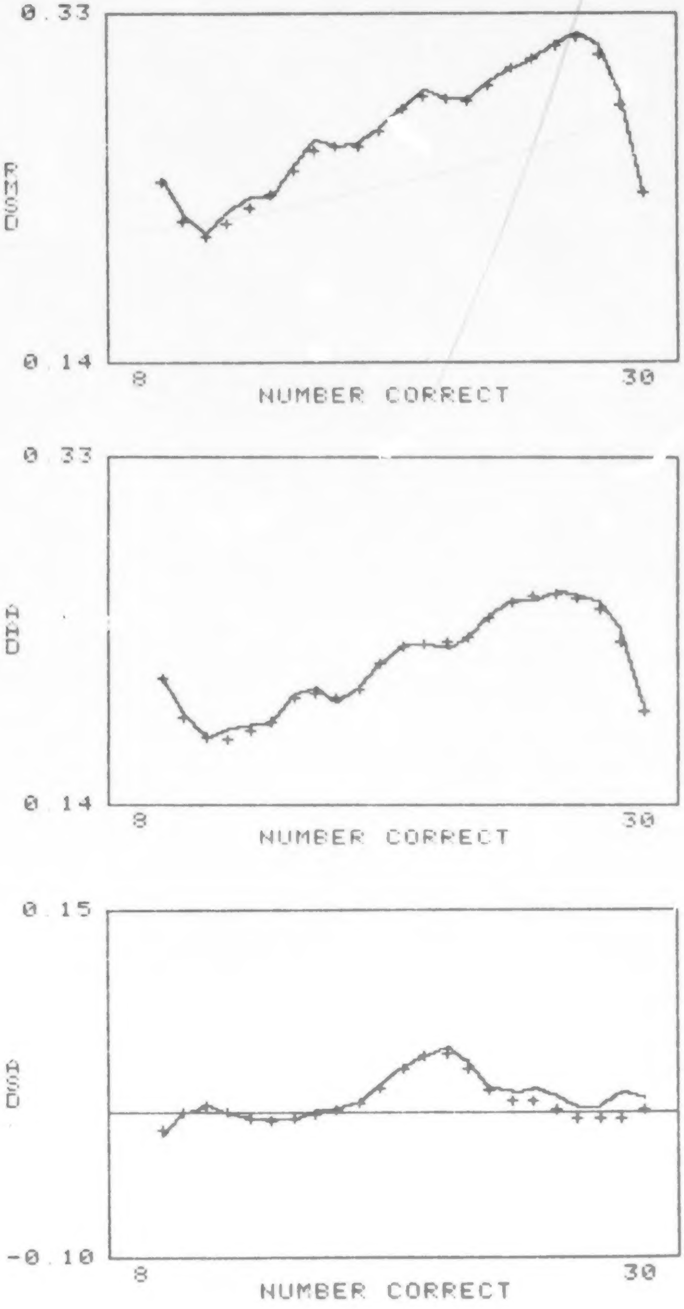


Figure B-23

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Presmoothed by 5-point moving weighted averages with root transformation

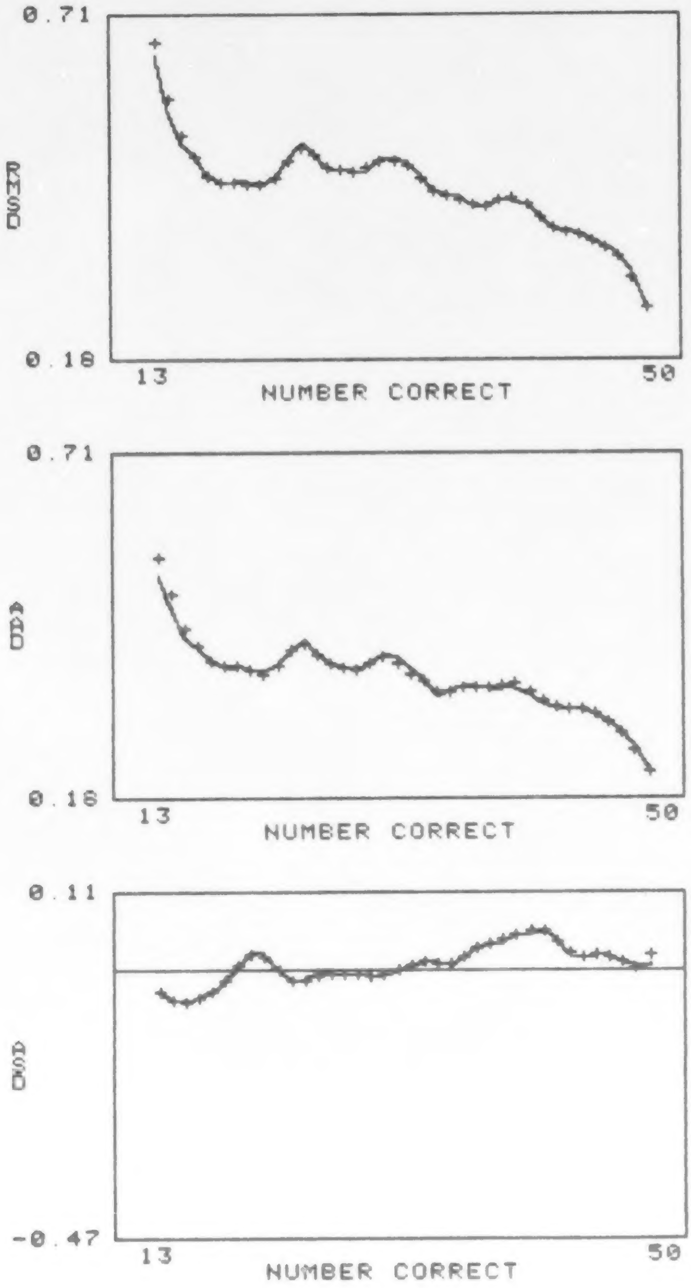


Figure B-24

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 20
Test Type: Operational
Smoothing: Presmoothed by 5-point moving weighted averages with root transformation

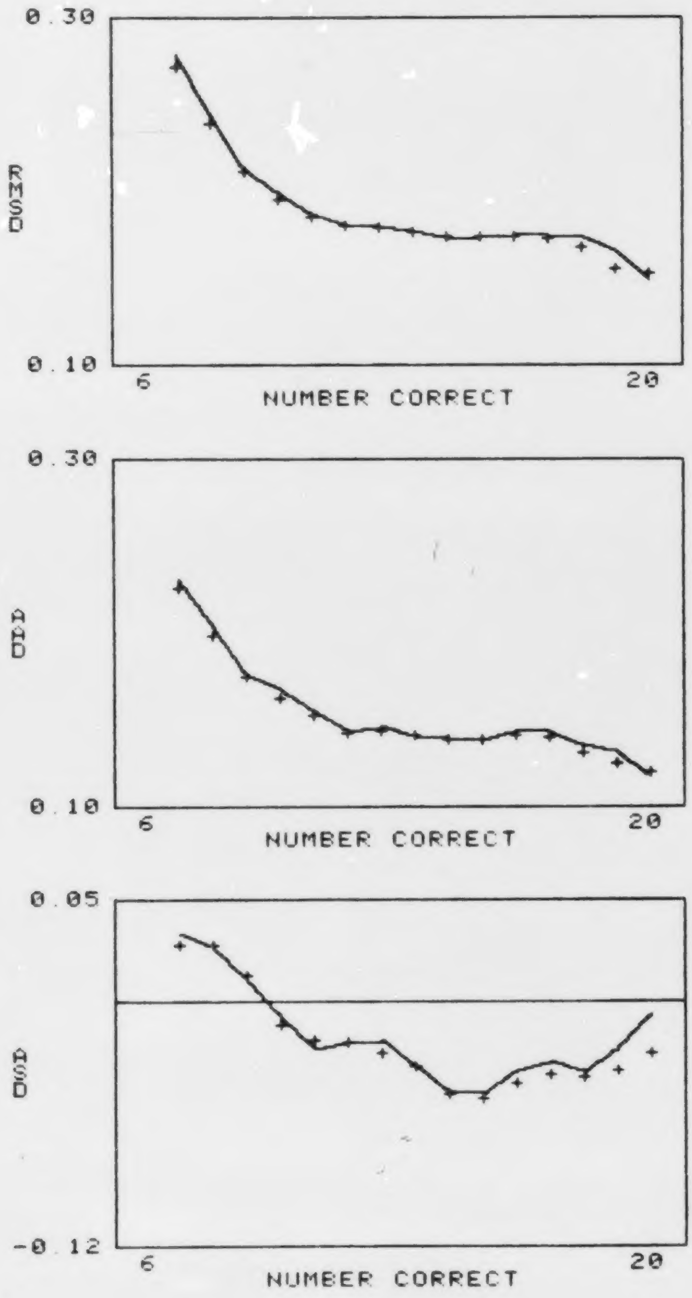


Figure B-25

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25
Test Type: Operational
Smoothing: Presmoothed by 5-point moving weighted averages with root transformation

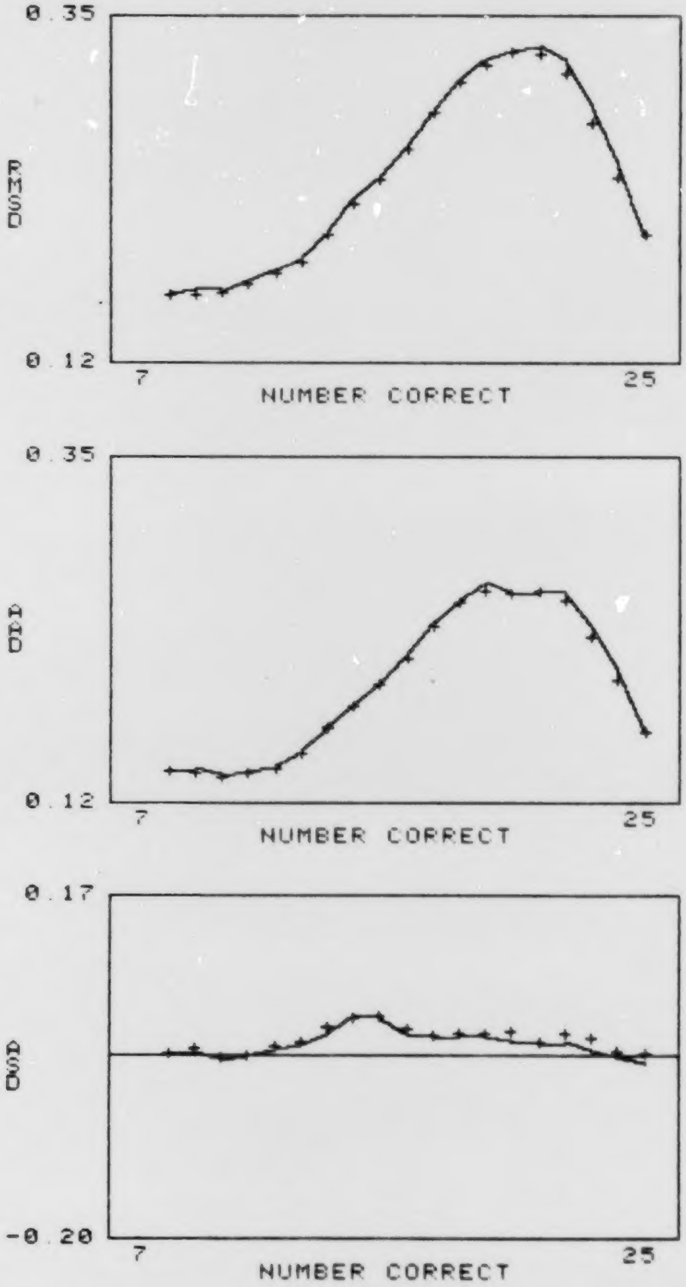


Figure B-26

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15
Test Type: Simulated
Smoothing: Presmoothed by 4253H Twice

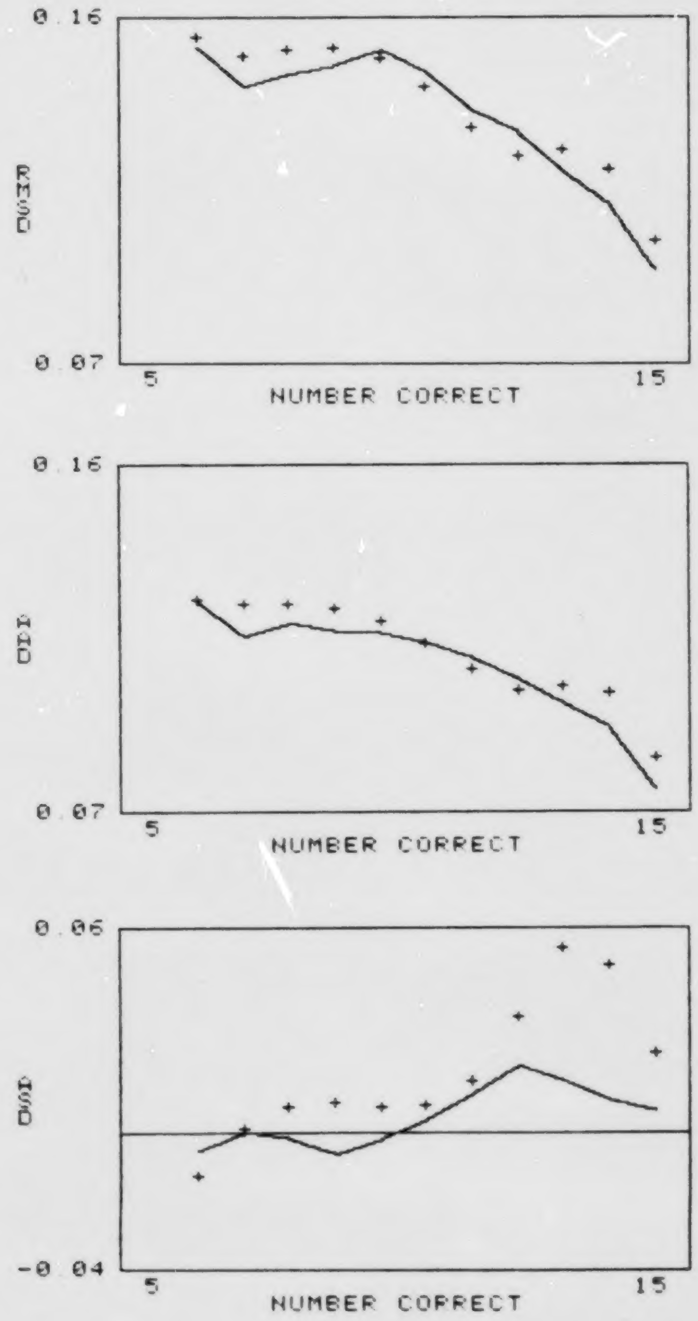


Figure B-27

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Presmoothed by 4253H Twice

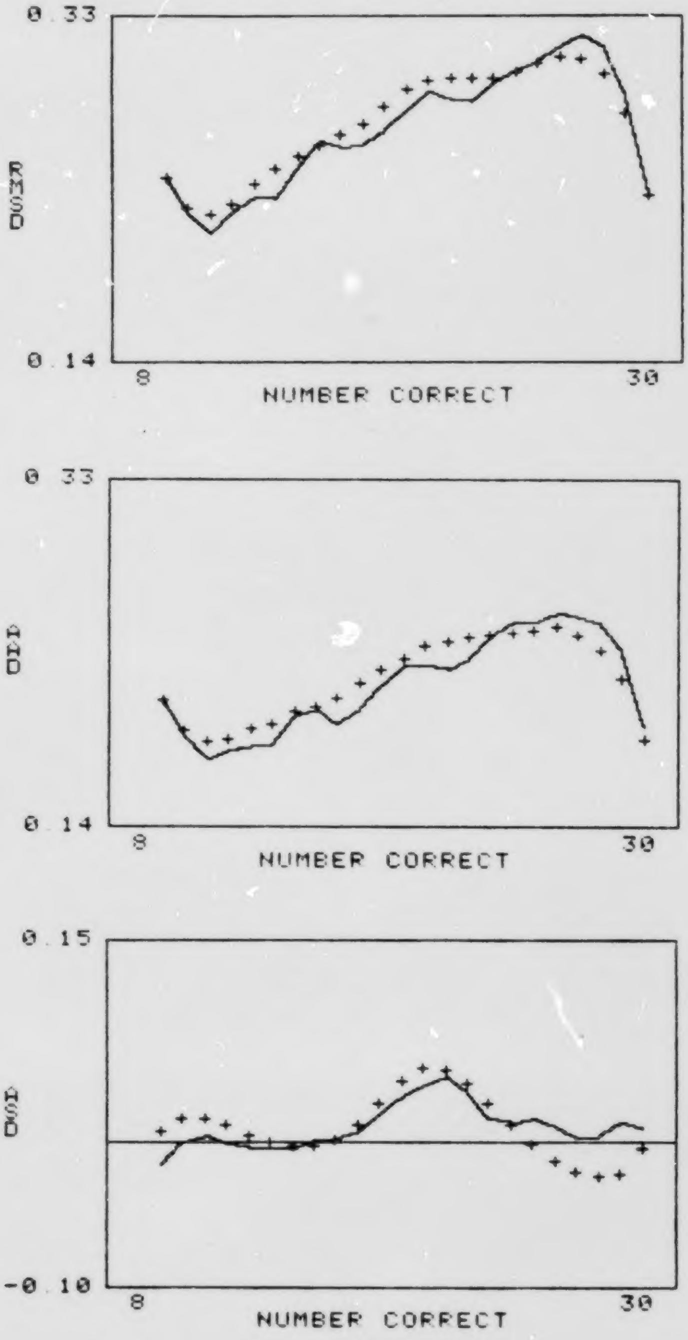


Figure B-28

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Presmoothed by 4253H Twice

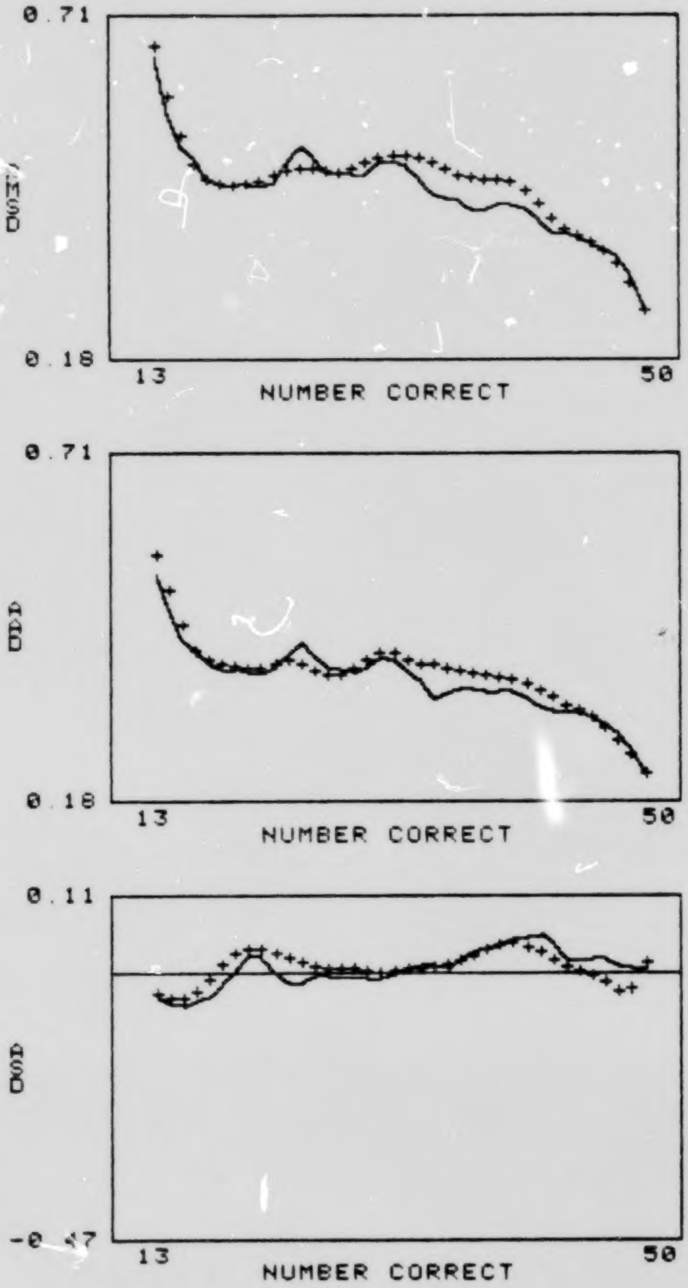


Figure B-29

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Presmoothed by 4253H Twice

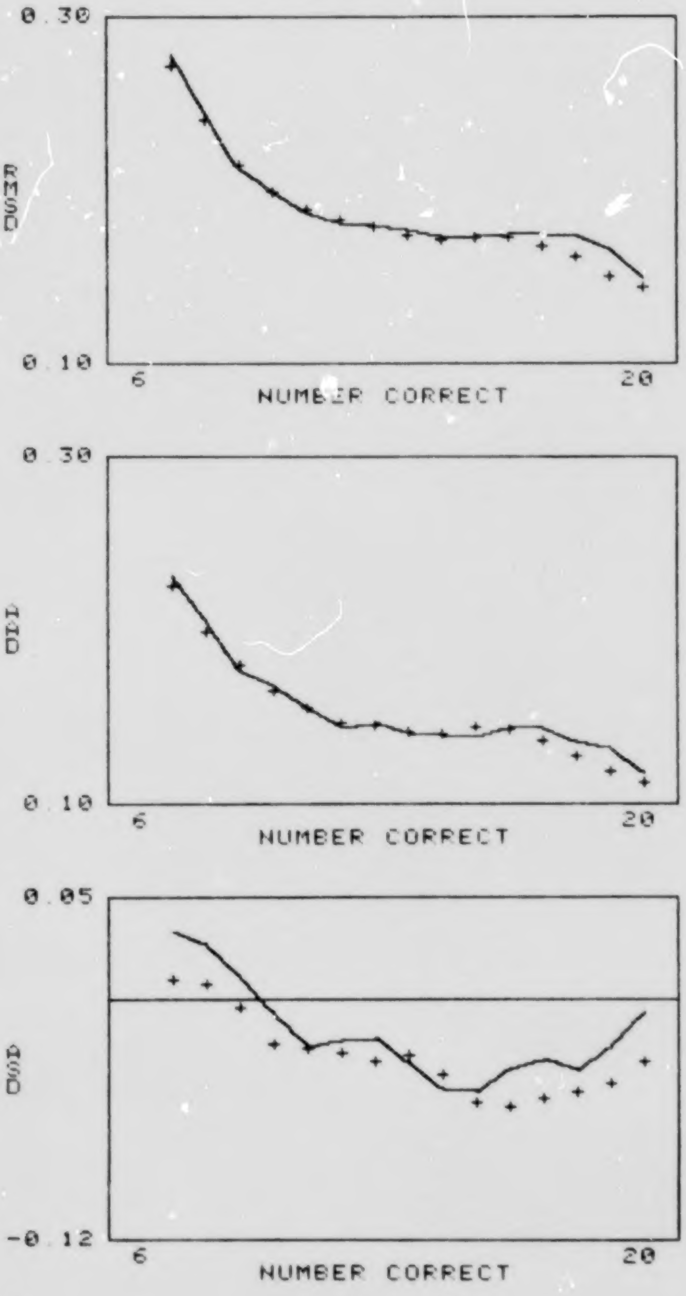


Figure B-30

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Presmoothed by 4253H Twice

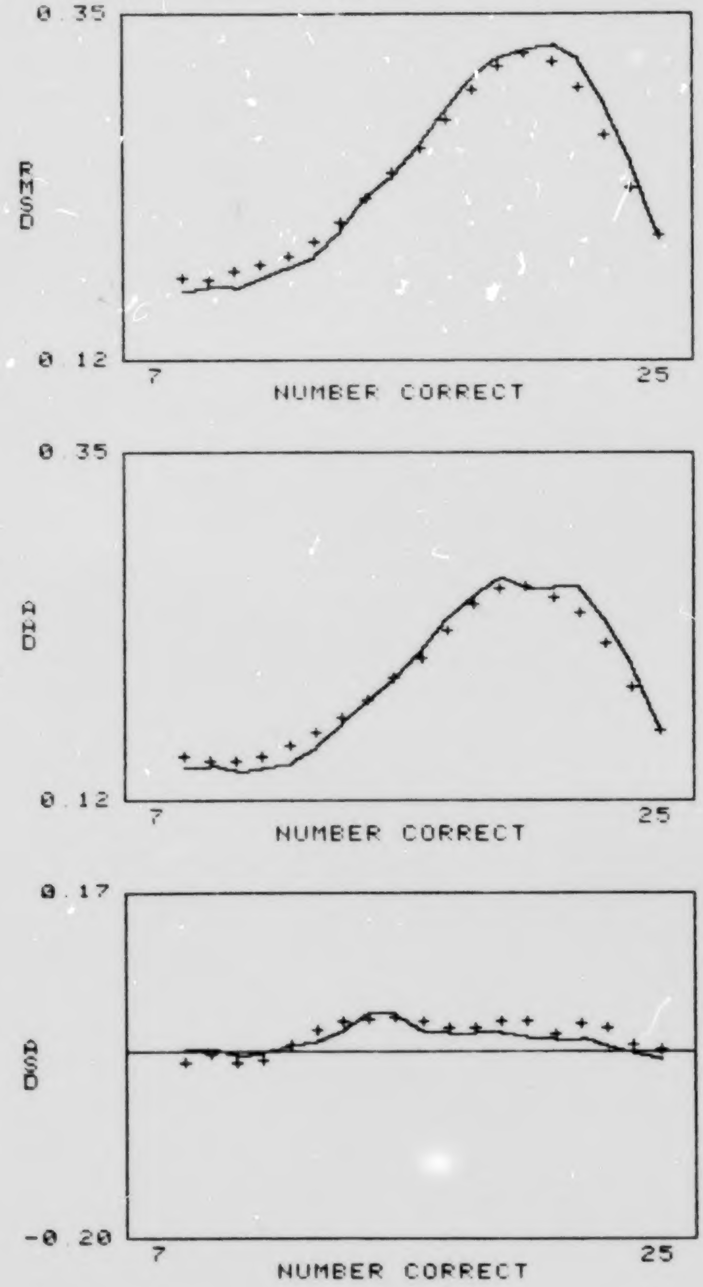


Figure B-31

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Presmoothed by negative hypergeometric

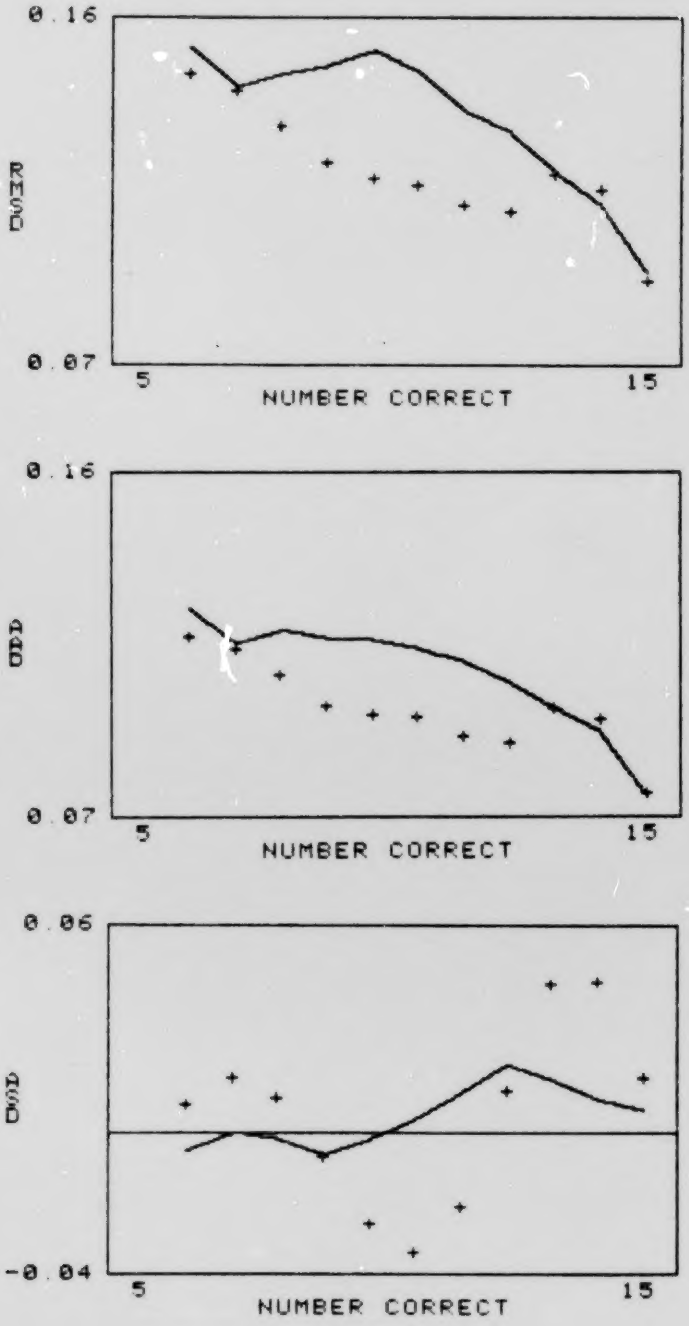


Figure B-32

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Presmoothed by negative hypergeometric

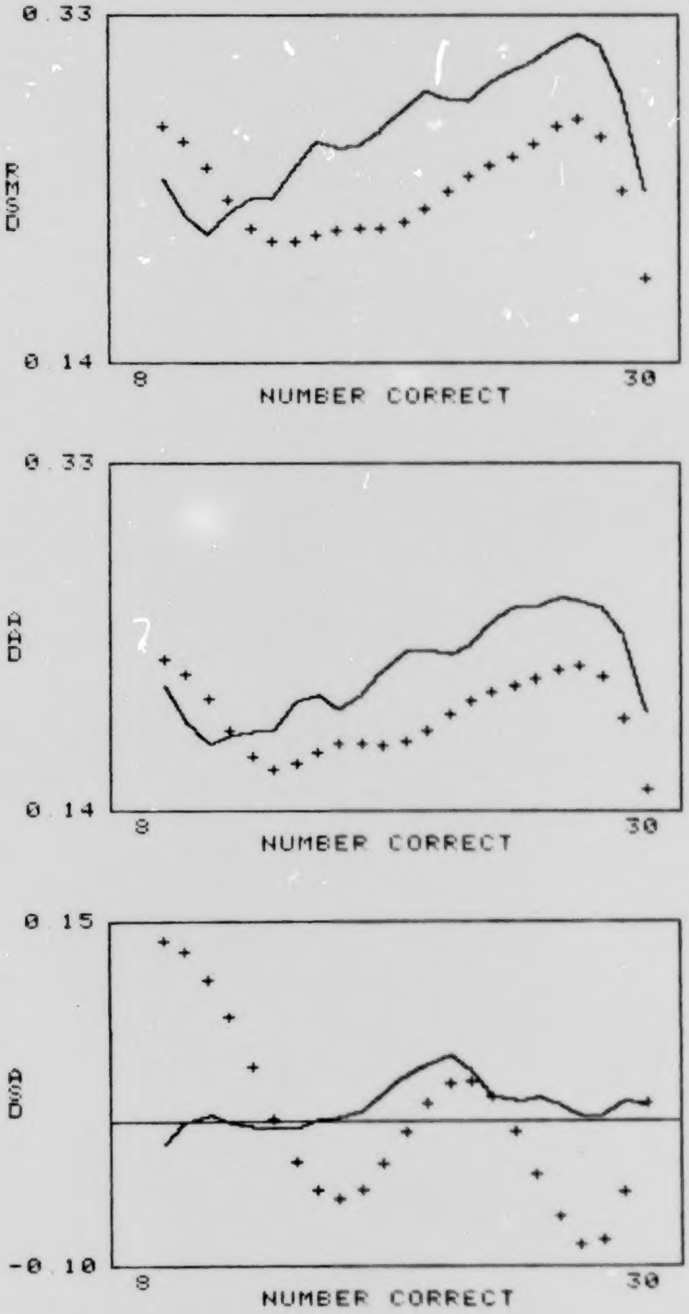


Figure B-33

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Presmoothed by negative hypergeometric

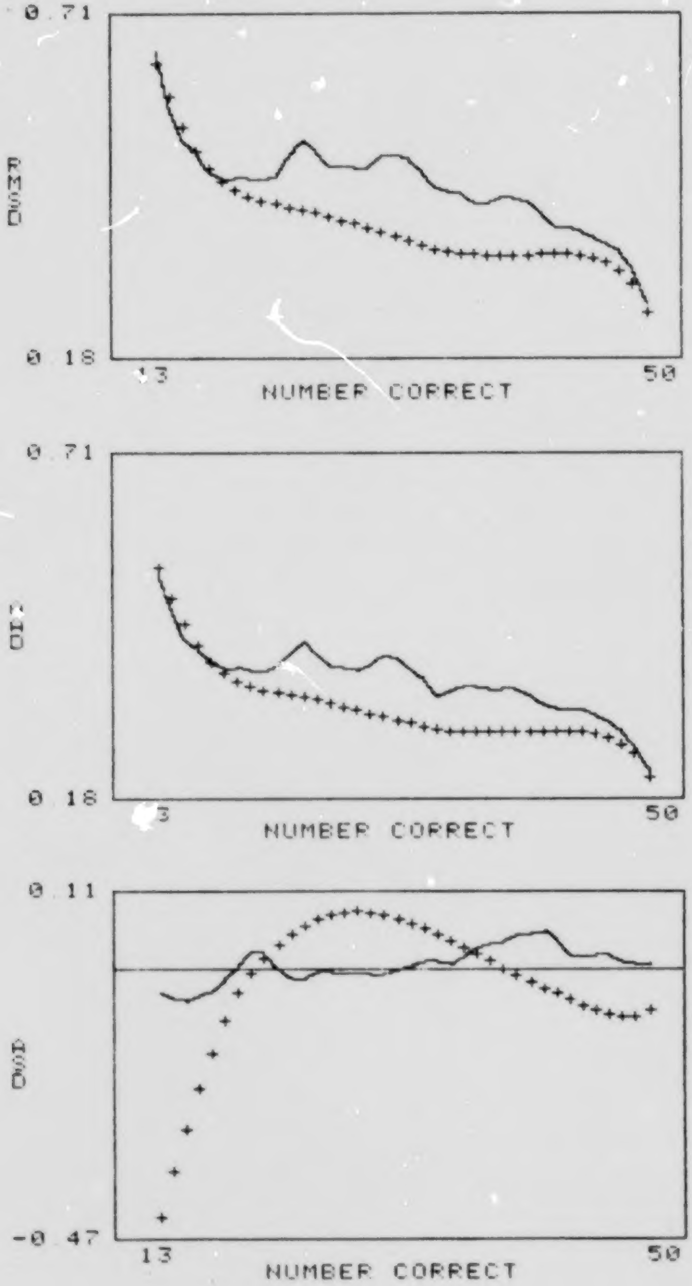


Figure B-34

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Presmoothed by negative hypergeometric

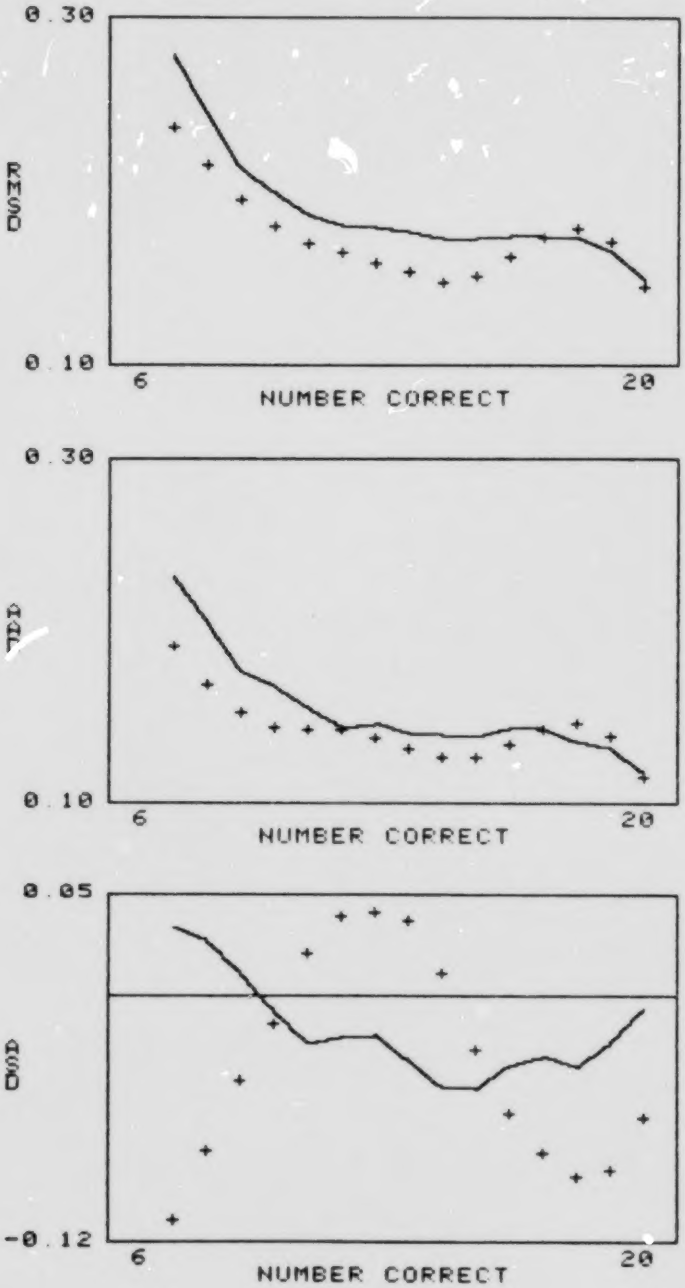
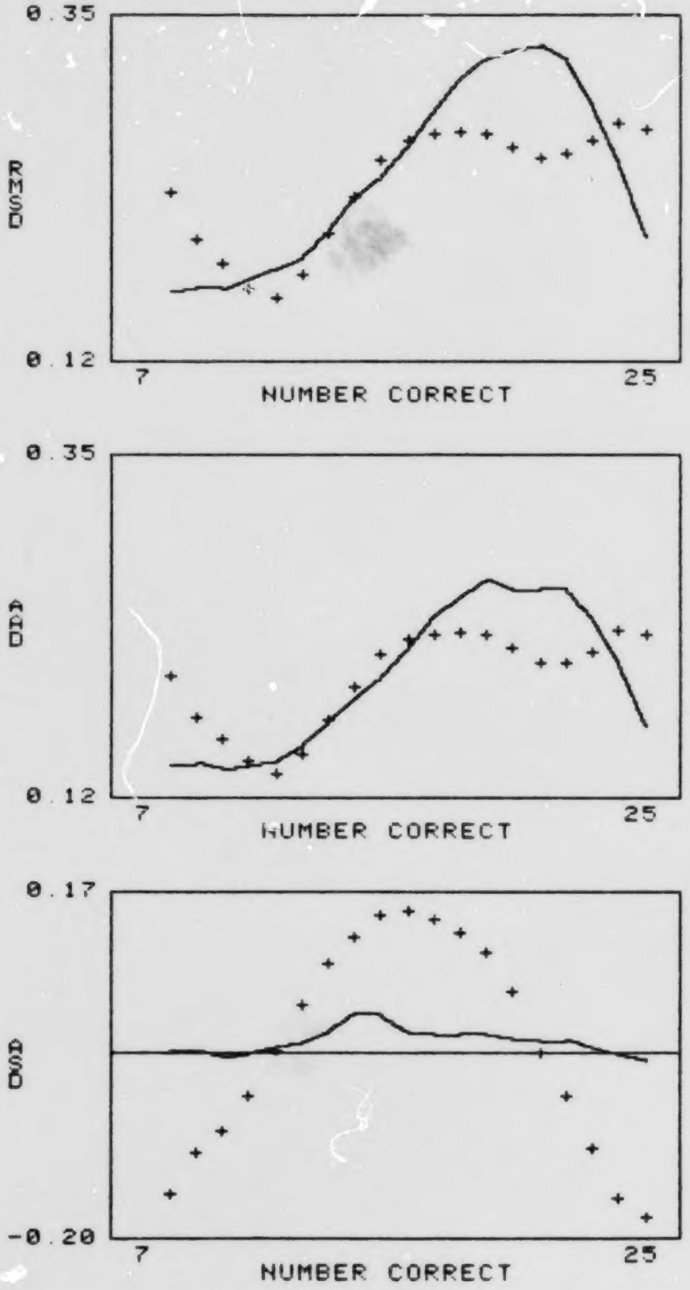


Figure B-35

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25
Test Type: Operational
Smoothing: Presmoothed by negative hypergeometric



108

Figure B-36

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Postsmoothed by linear regression

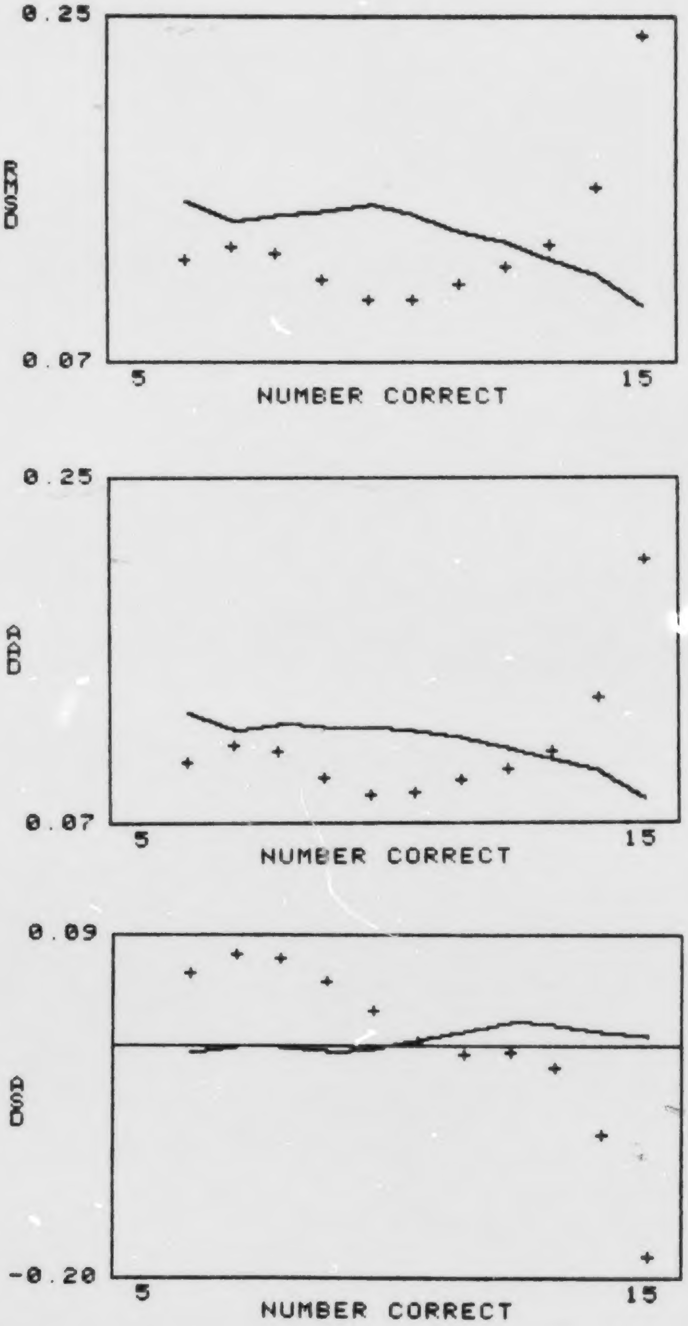


Figure B-37

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Postsmoothed by linear regression

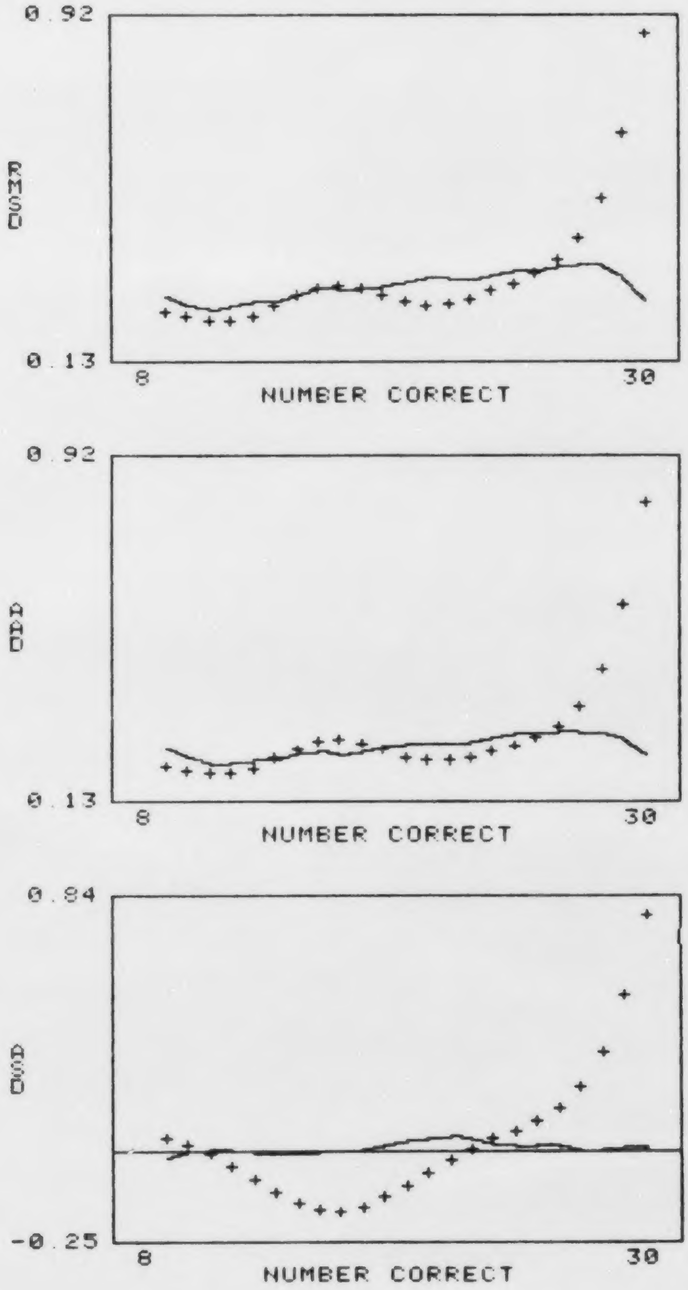


Figure B-38

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Postsmoothed by linear regression

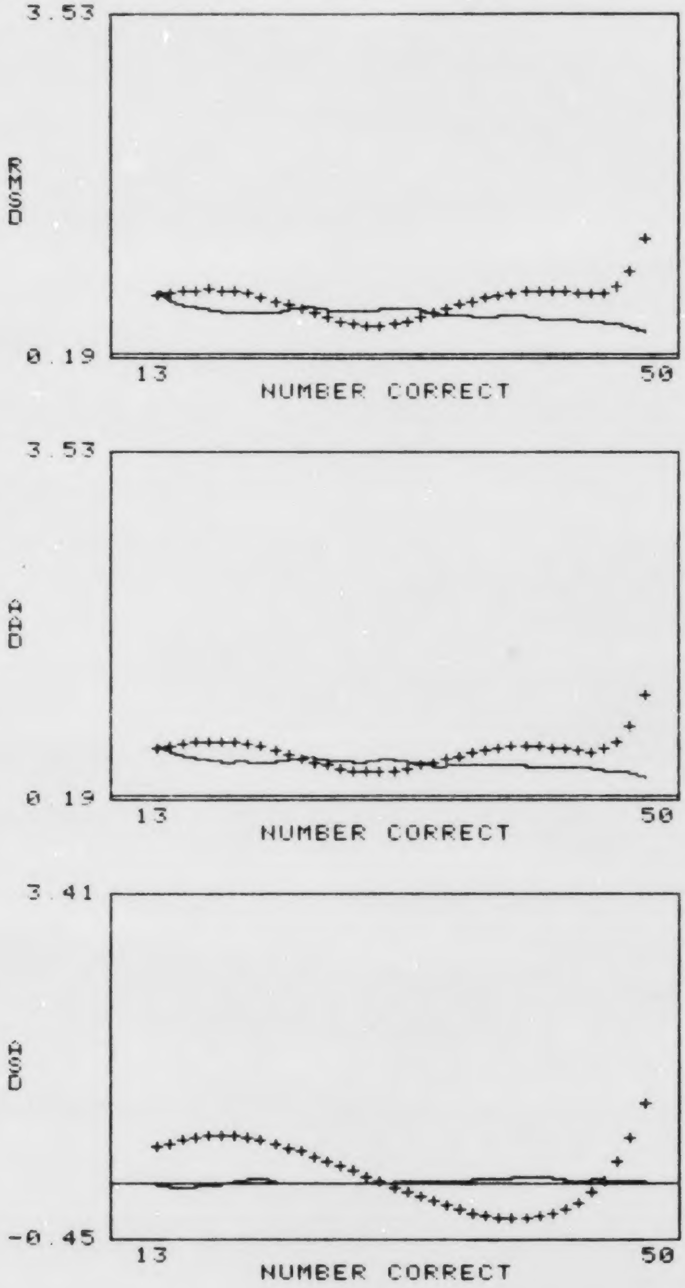


Figure B-39

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Postsmoothed by linear regression

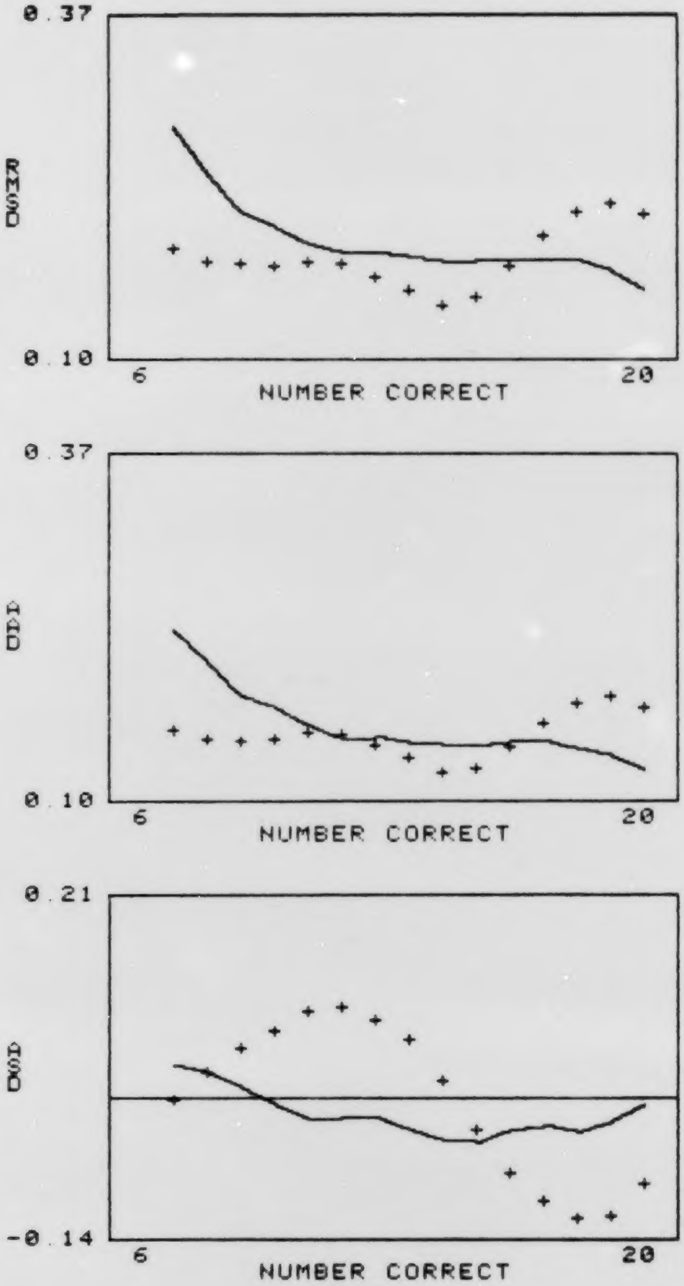


Figure B-40

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25

Test Type: Operational

Smoothing: Postsmoothed by linear regression

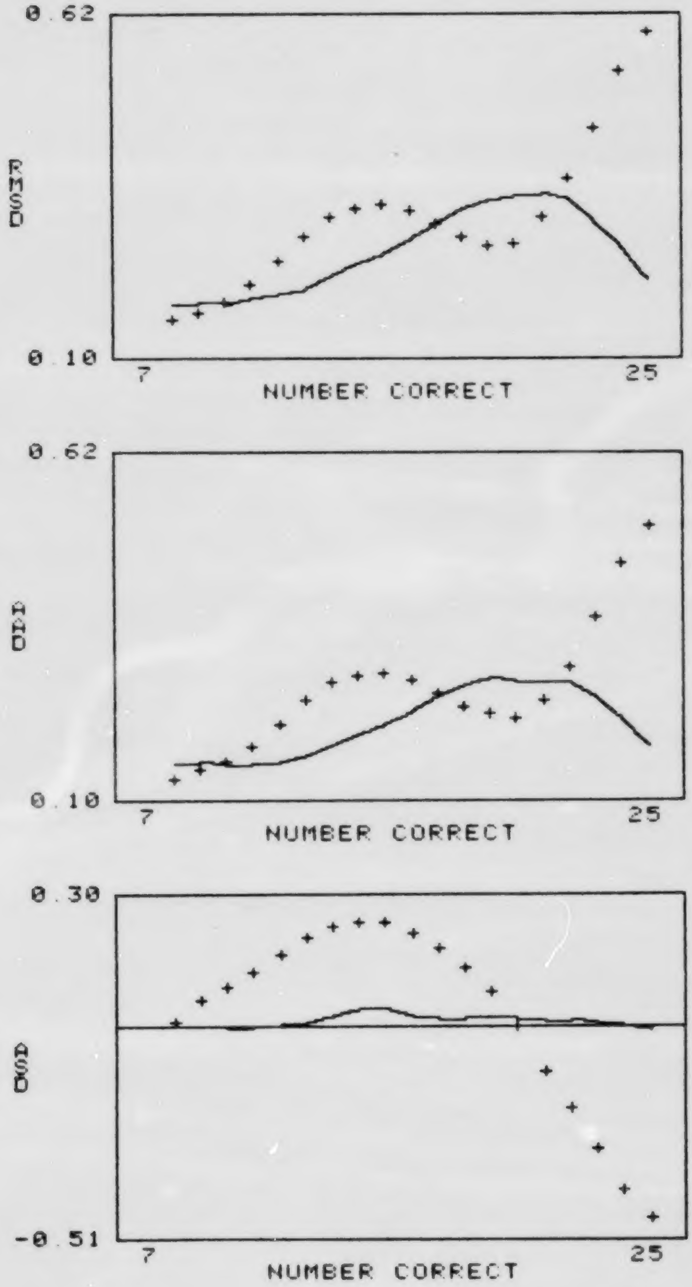


Figure B-41

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15
Test Type: Simulated
Smoothing: Postsmoothed by quadratic regression

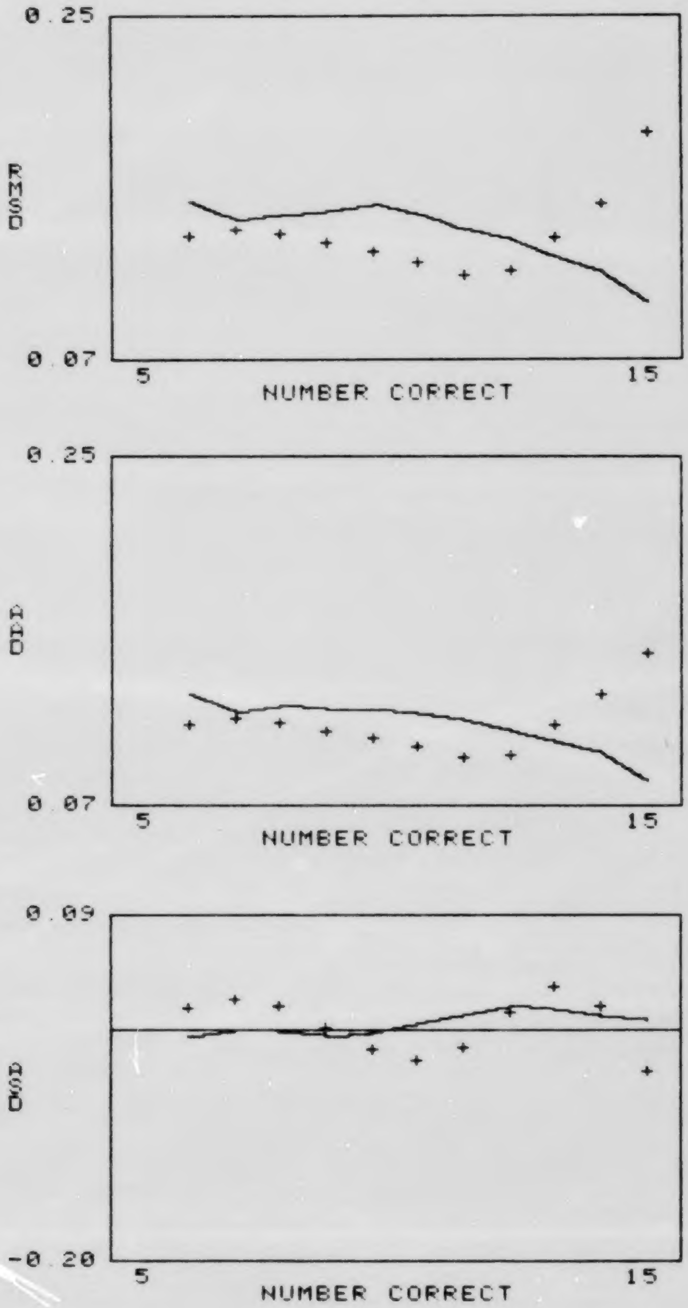


Figure B-42

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Postsmoothed by quadratic regression

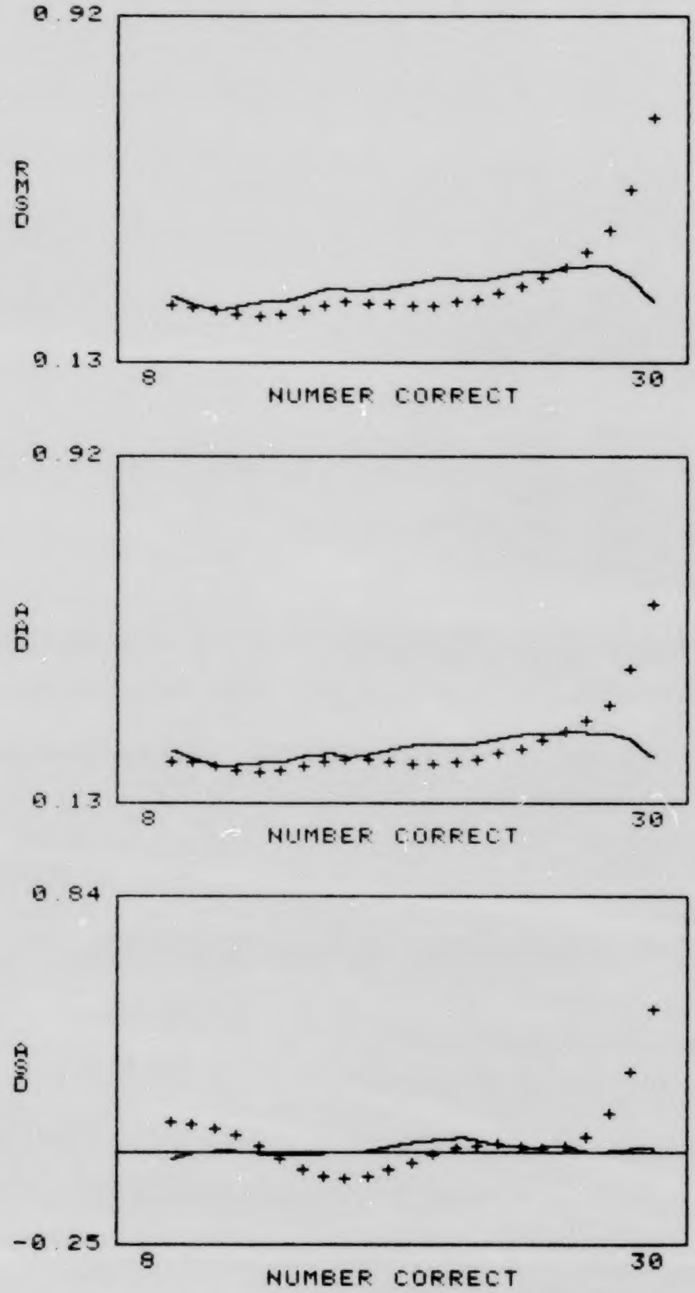


Figure B-43

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Postsmoothed by quadratic regression

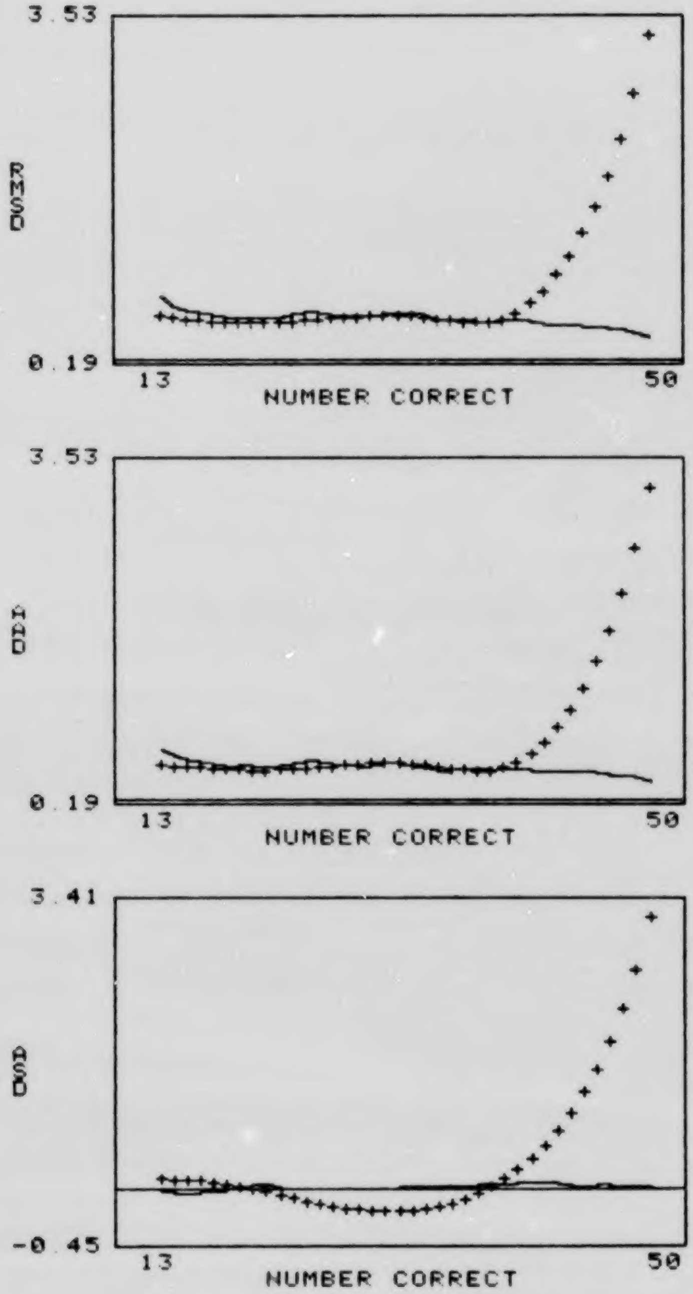


Figure B-44

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Postsmoothed by quadratic regression

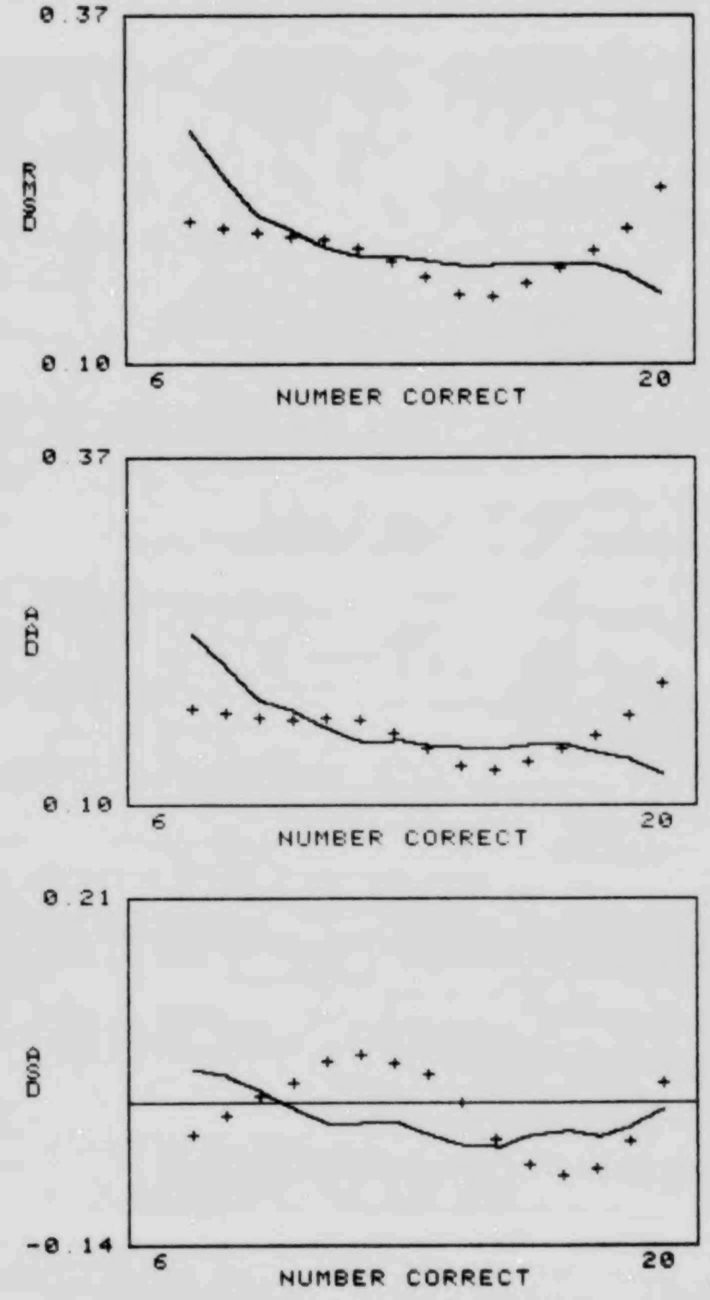


Figure B-45

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25

Test Type: Operational

Smoothing: Postsmoothed by quadratic regression

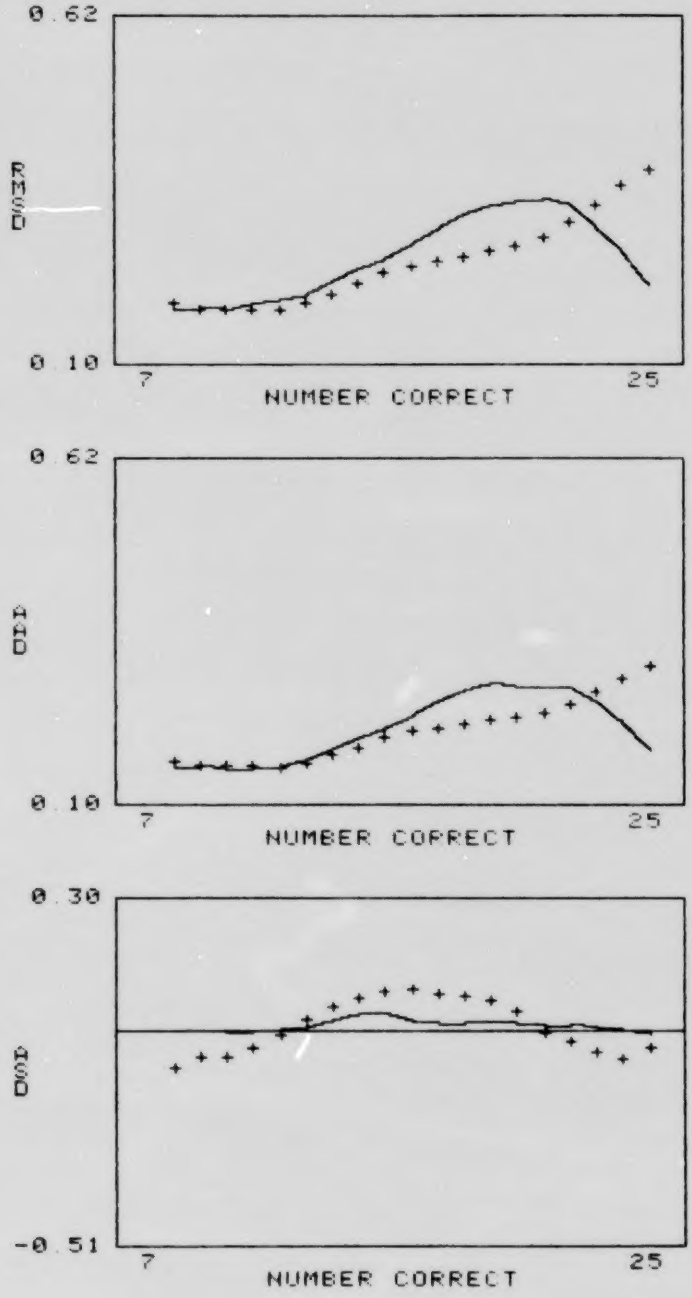


Figure B-46

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15
Test Type: Simulated
Smoothing: Postsmoothed by cubic regression

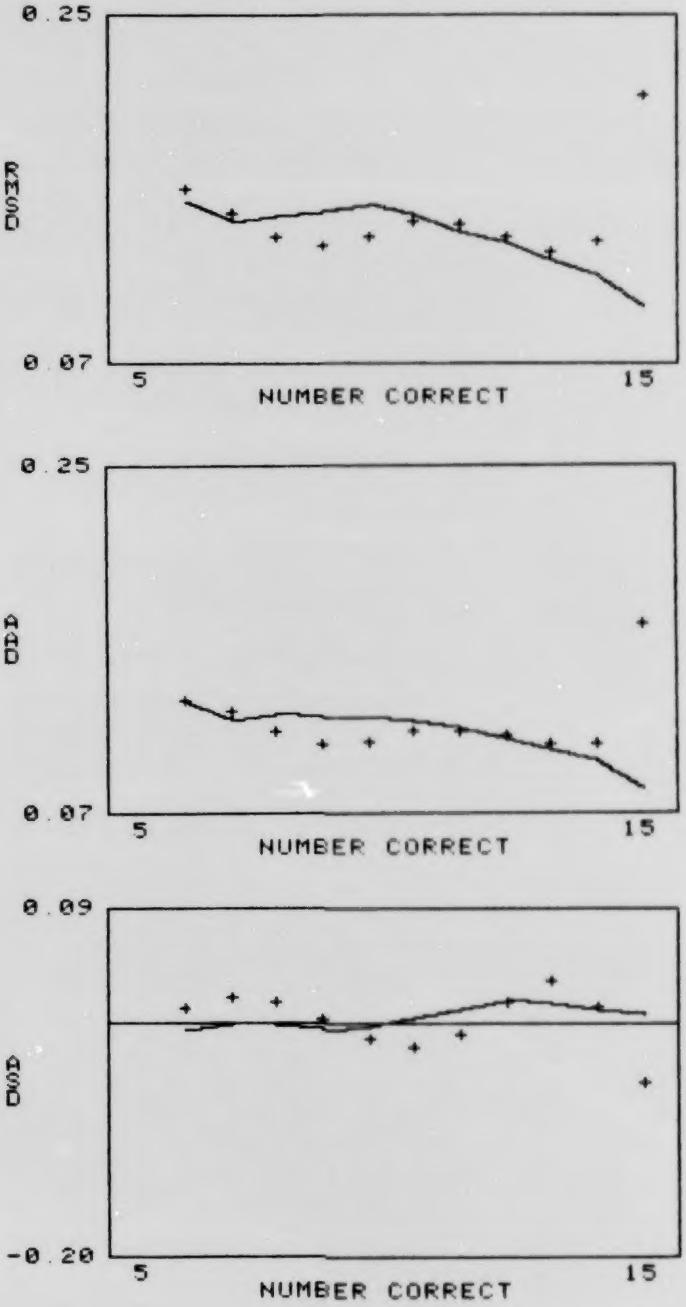


Figure B-47

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Postsmoothed by cubic regression

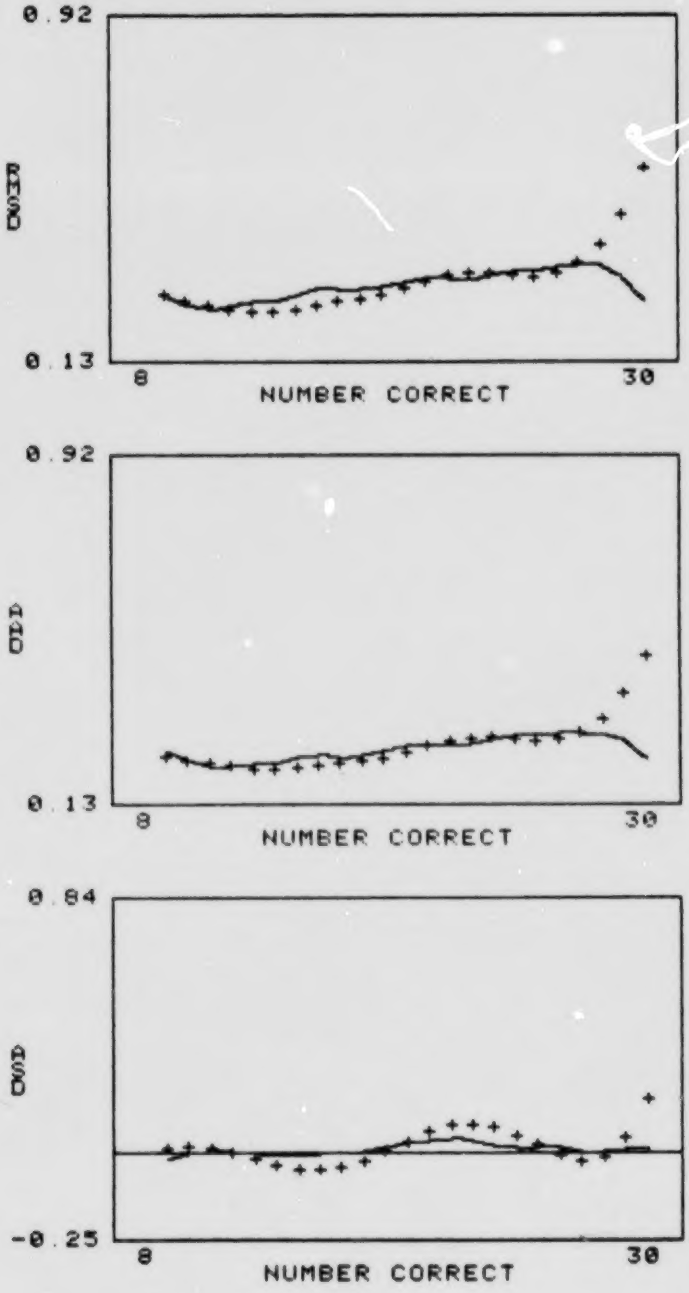


Figure B-48

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50

Test Type: Simulated

Smoothing: Postsmoothed by cubic regression

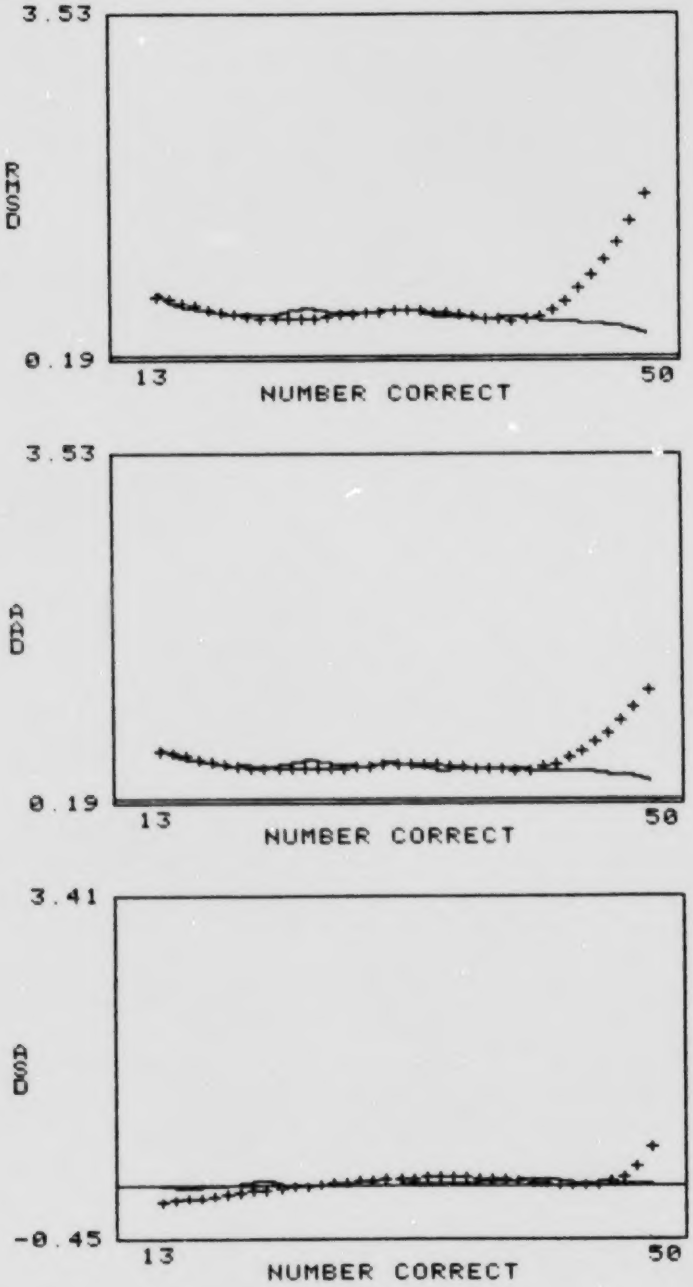


Figure B-49

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Postsmoothed by cubic regression

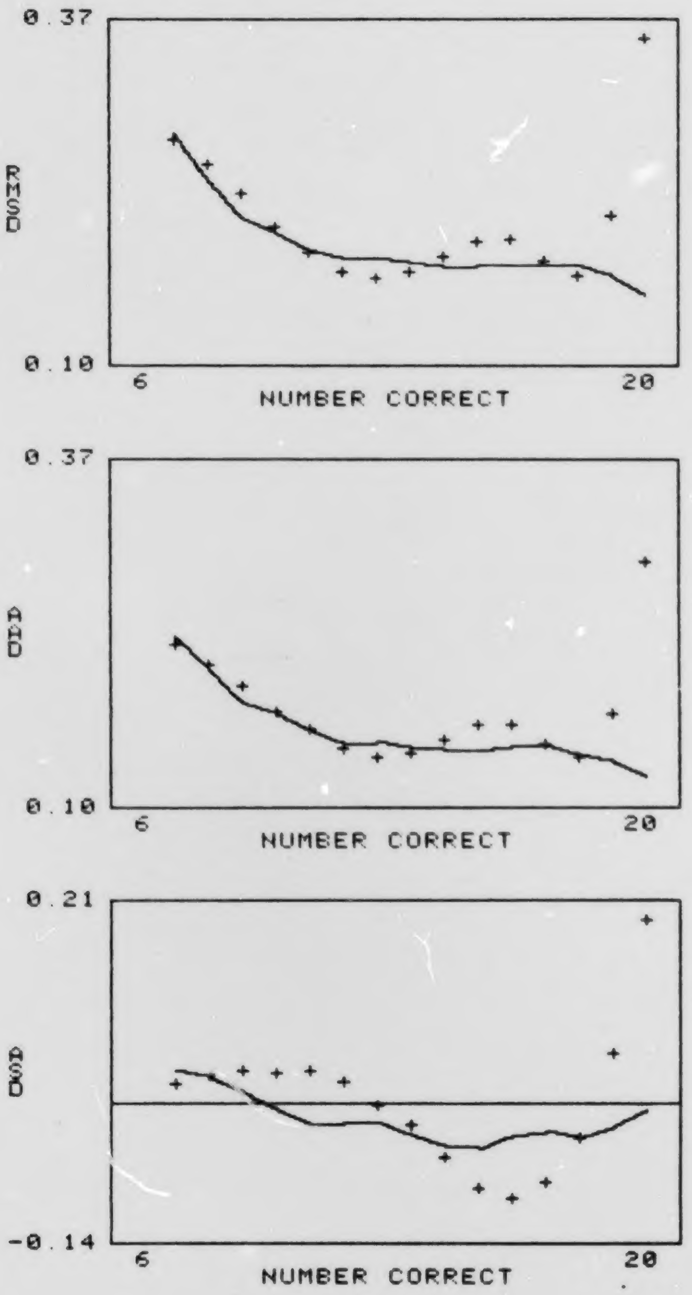


Figure B-50

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Postsmoothed by cubic regression

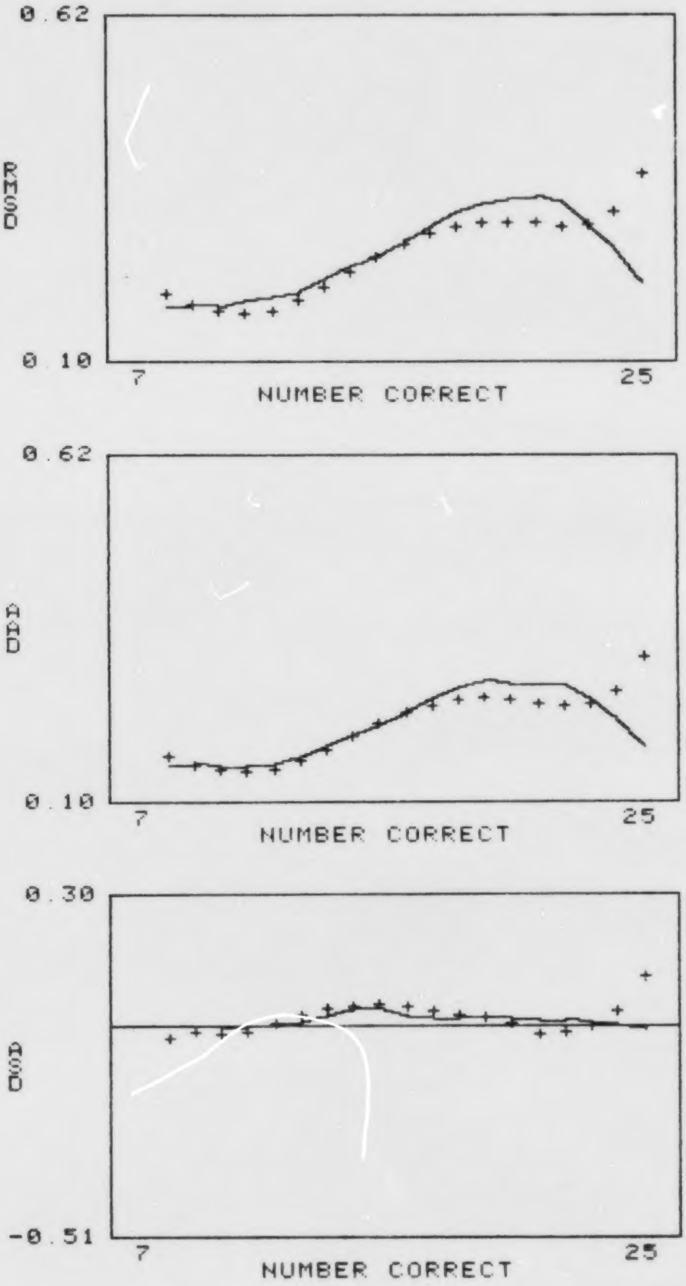


Figure B-51

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Postsmoothed by orthogonal regression

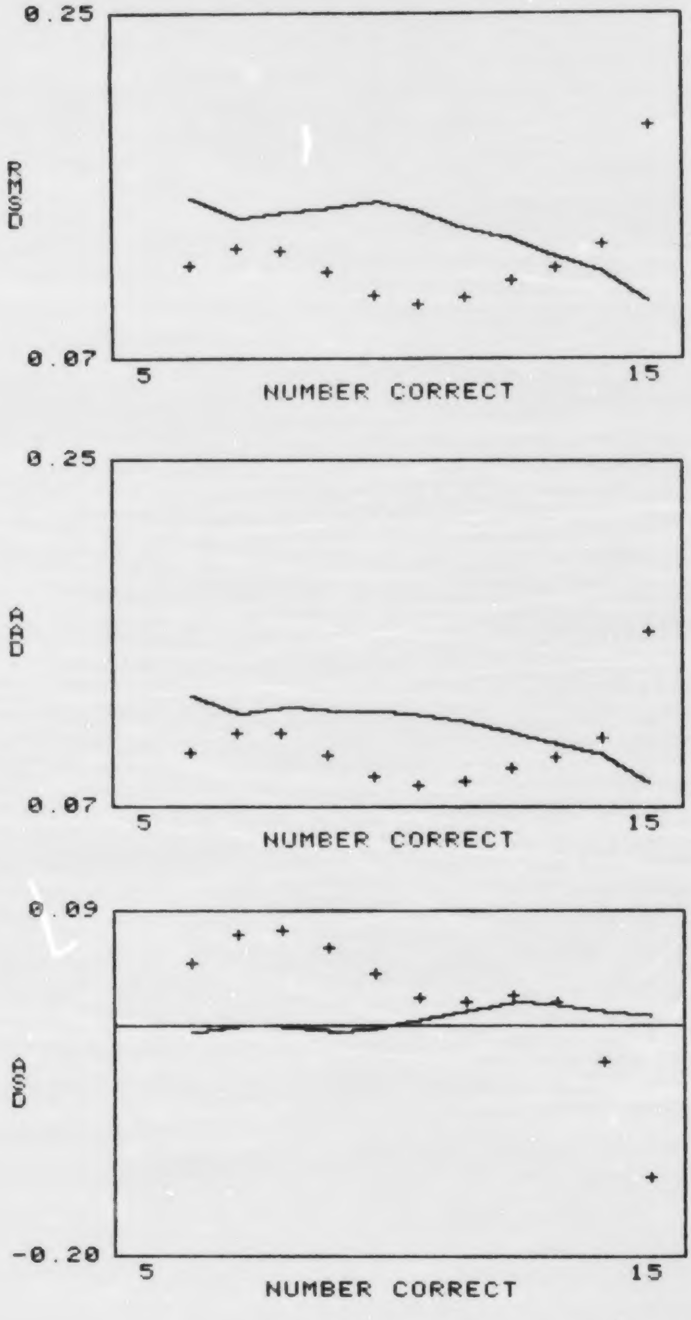


Figure B-52

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Postsmoothed by orthogonal regression

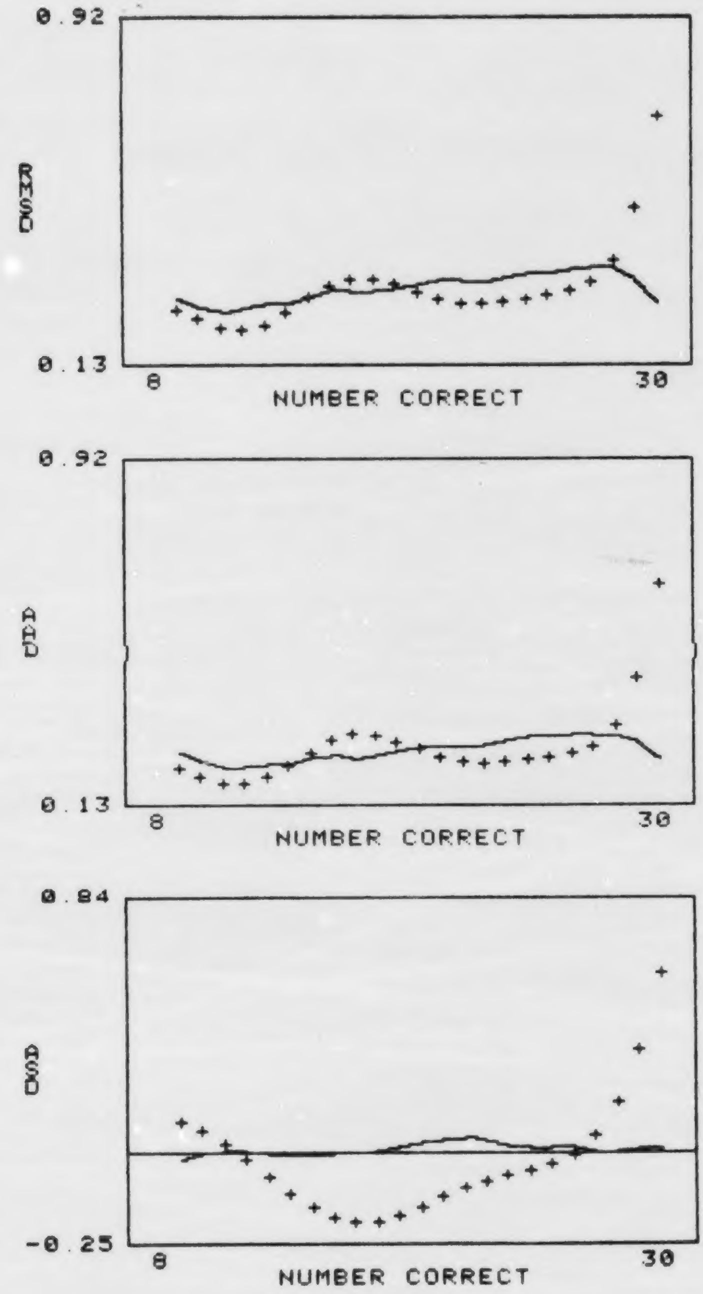


Figure B-53

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 50
Test Type: Simulated
Smoothing: Postsmoothed by orthogonal regression

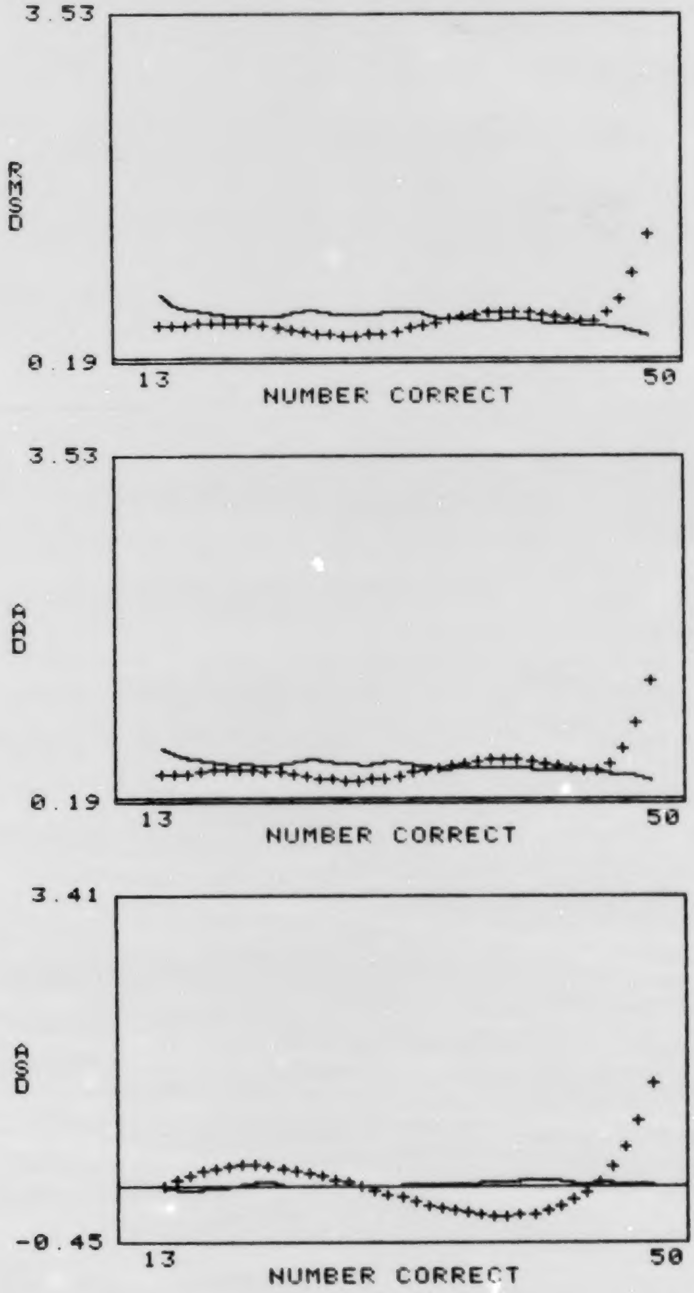


Figure B-54

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 20
Test Type: Operational
Smoothing: Postsmoothed by orthogonal regression

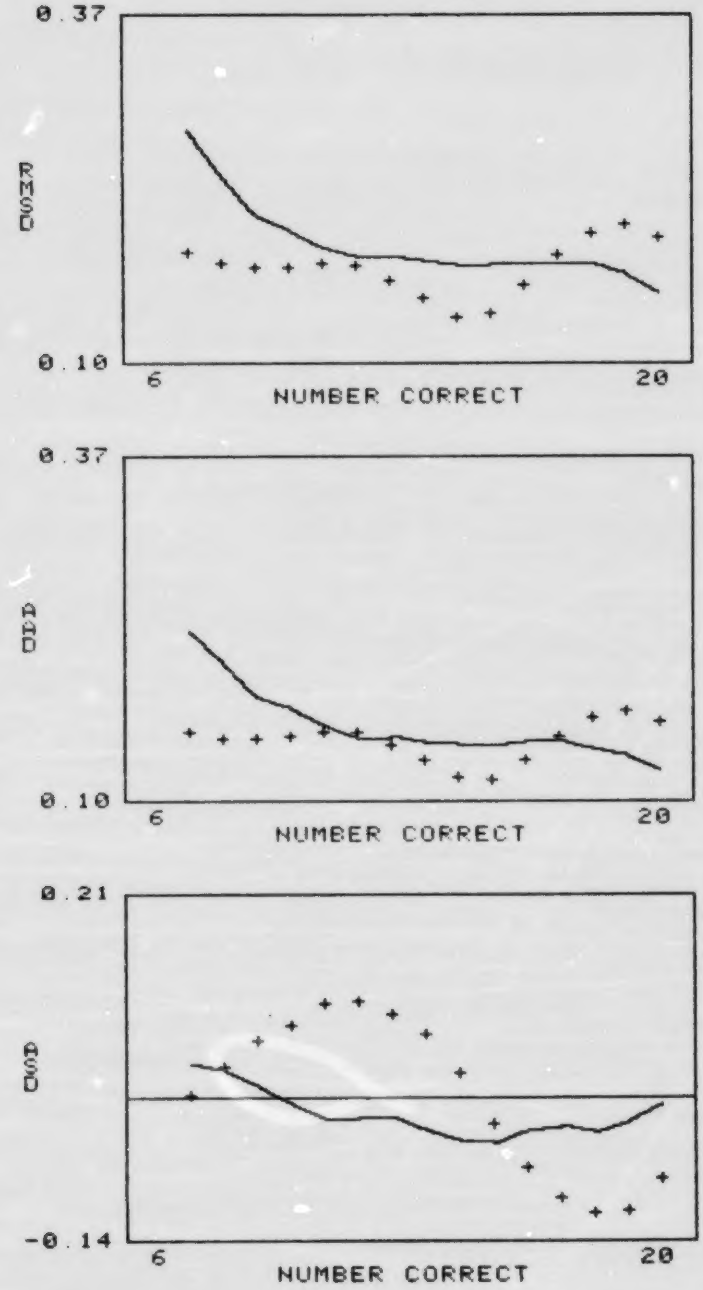


Figure B-55

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Postsmoothed by orthogonal regression

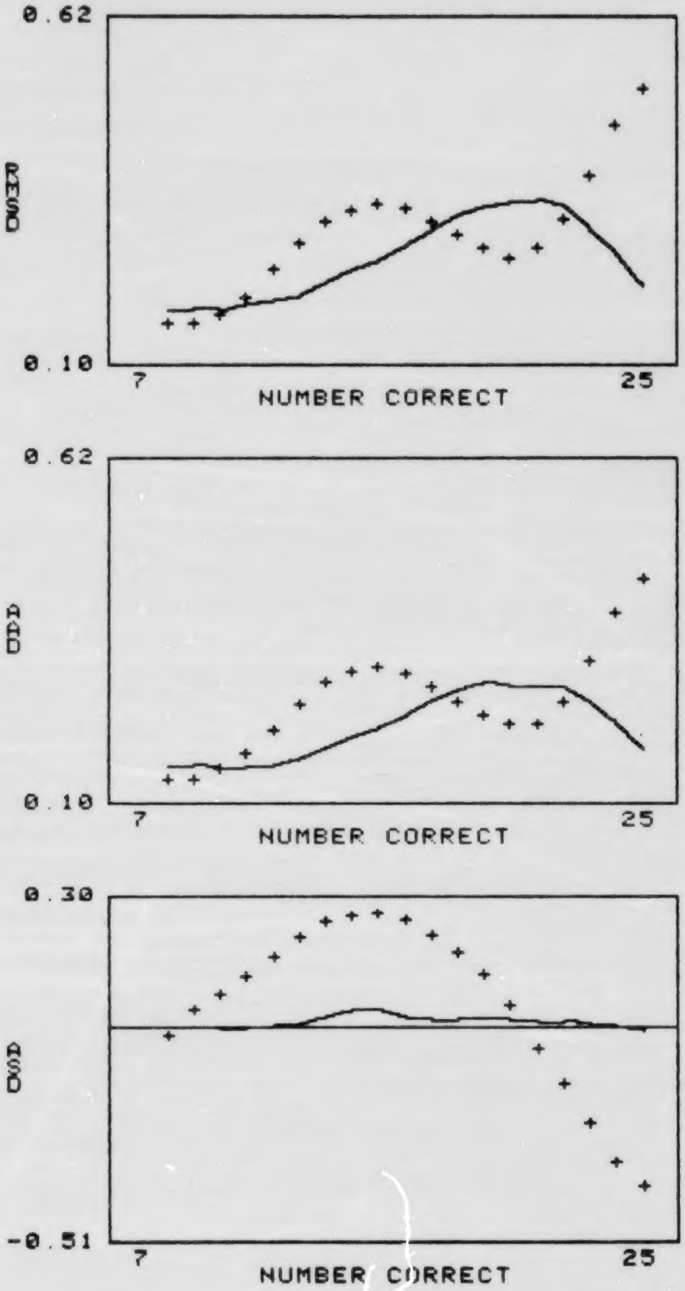


Figure B-56

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Postsmoothed by logistic ogive

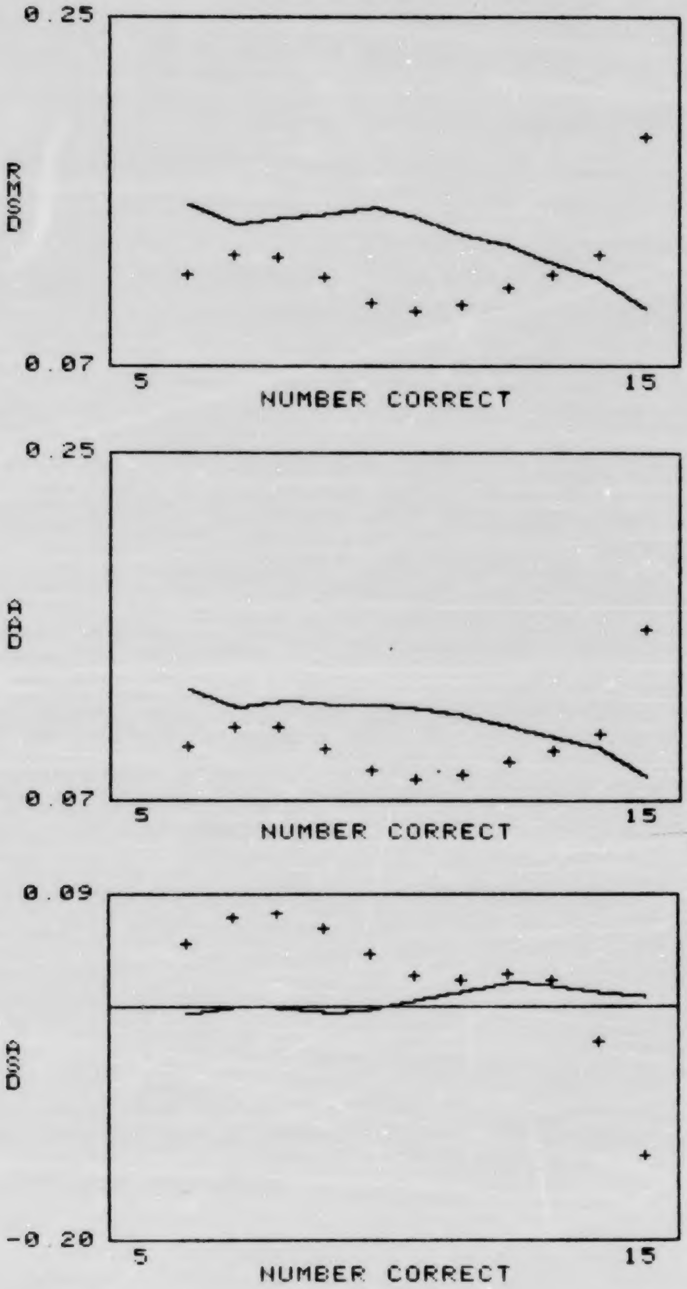


Figure B-57

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Postsmoothed by logistic ogive

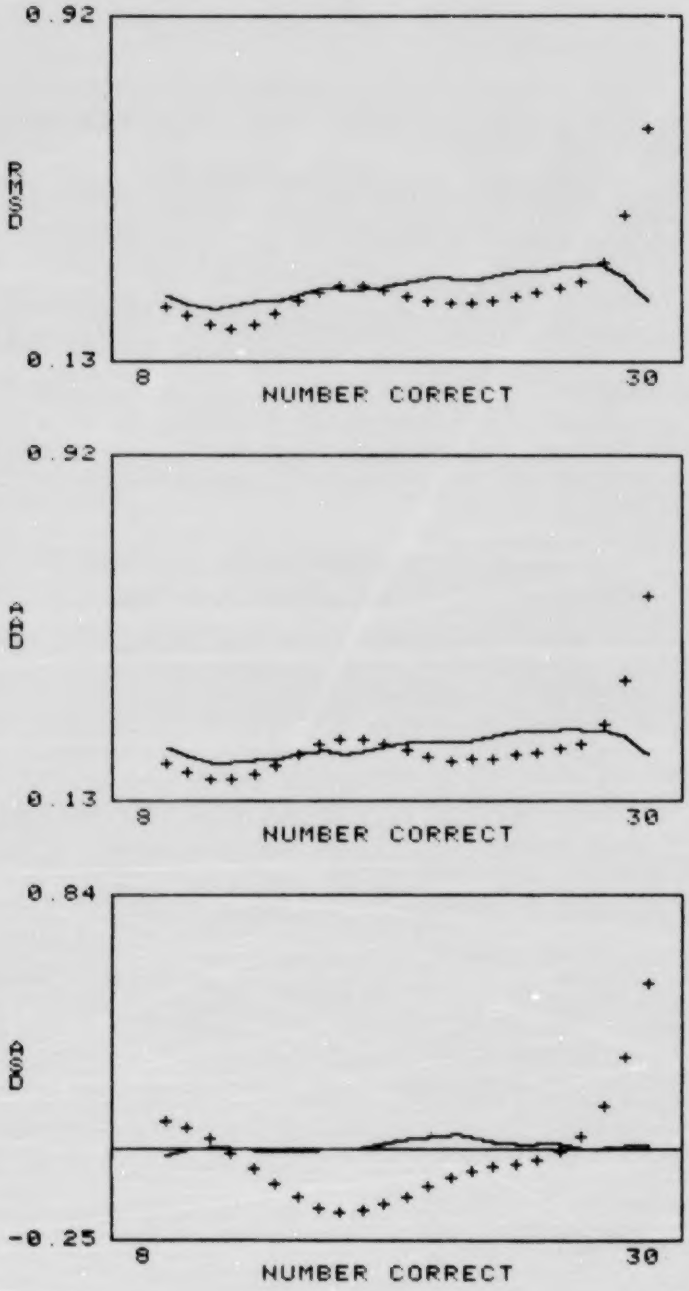


Figure B-58

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Postsmoothed by logistic ogive

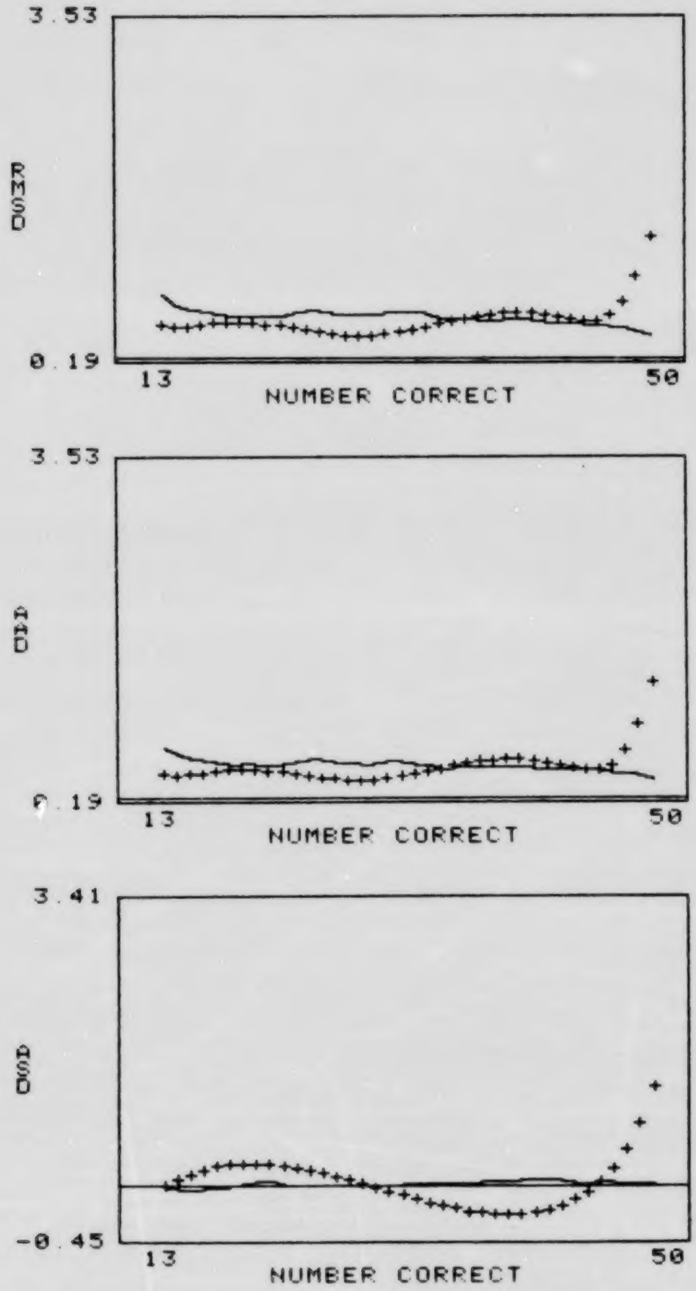


Figure B-59

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 20
Test Type: Operational
Smoothing: Postsmoothed by logistic ogive

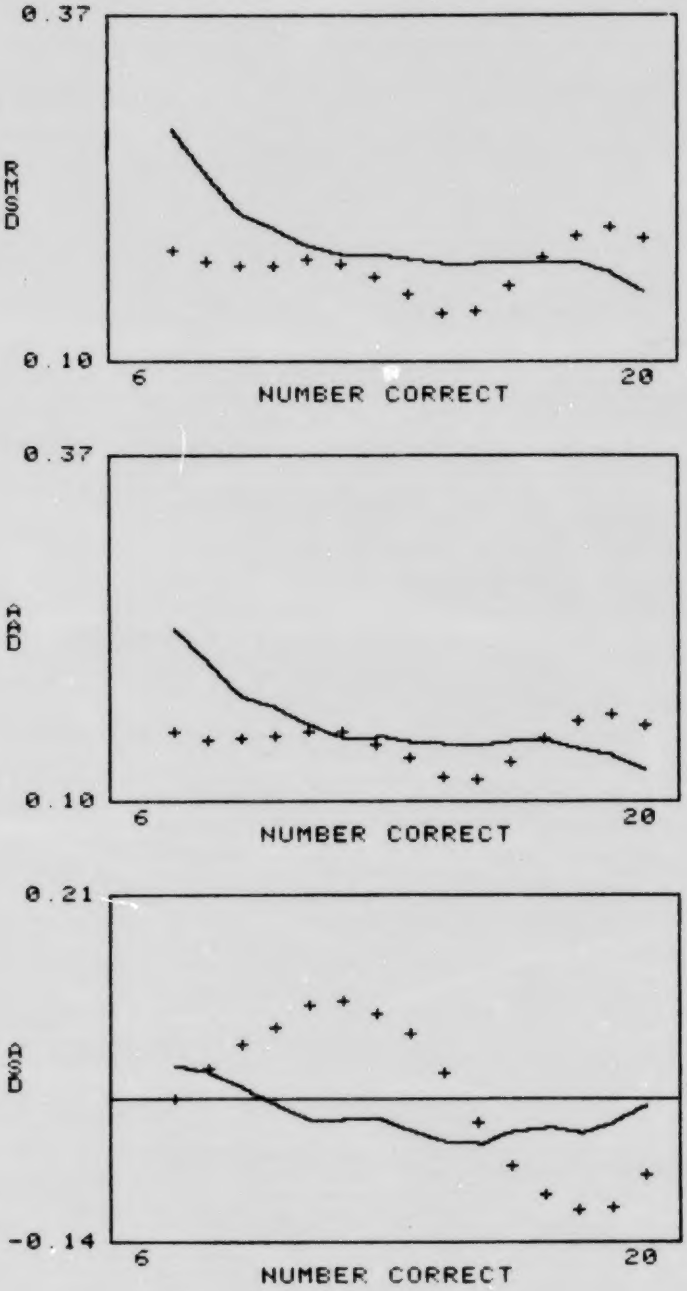


Figure B-60

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25
Test Type: Operational
Smoothing: Postsmoothed by logistic ogive

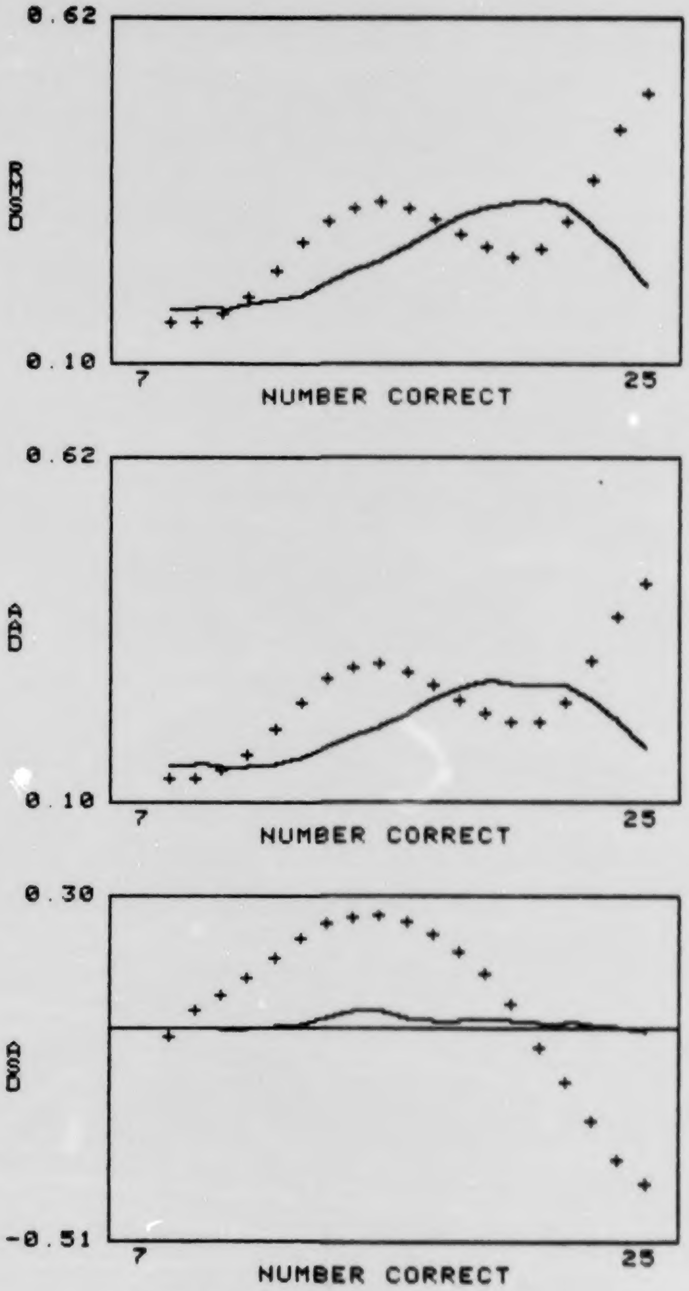


Figure B-61

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Postsmoothed by cubic smoothing splines

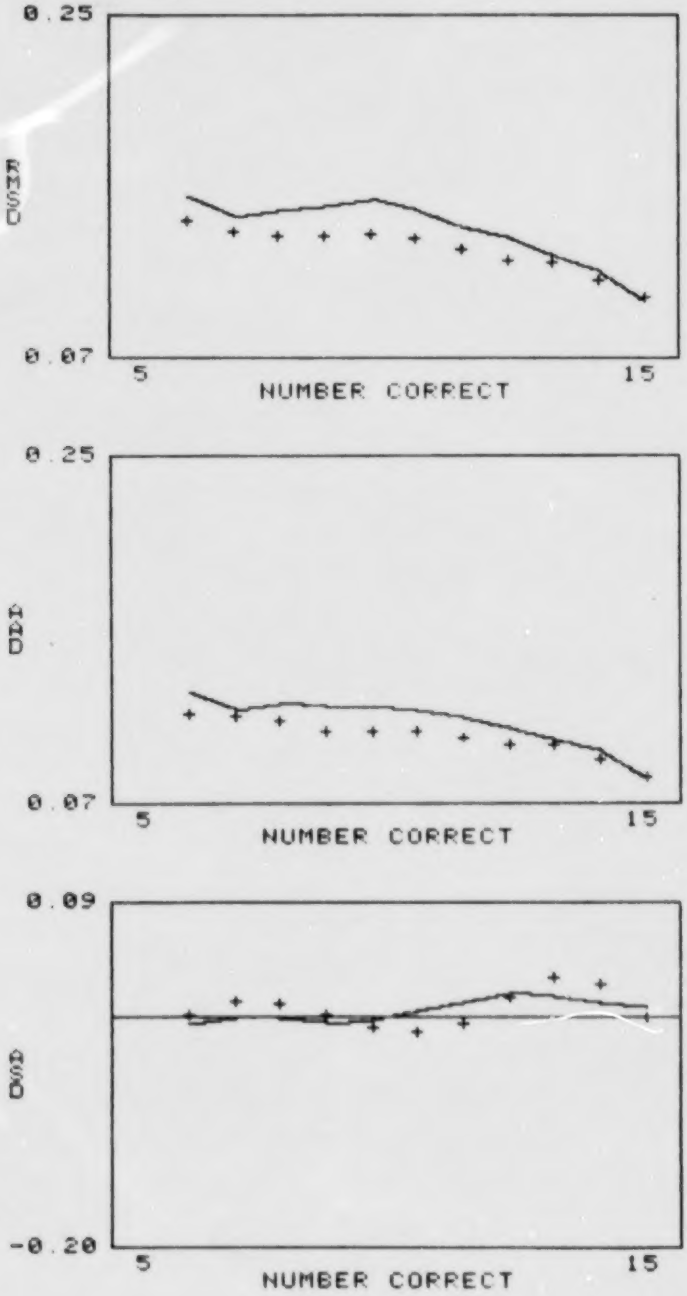


Figure B-62

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Postsmoothed by cubic smoothing splines

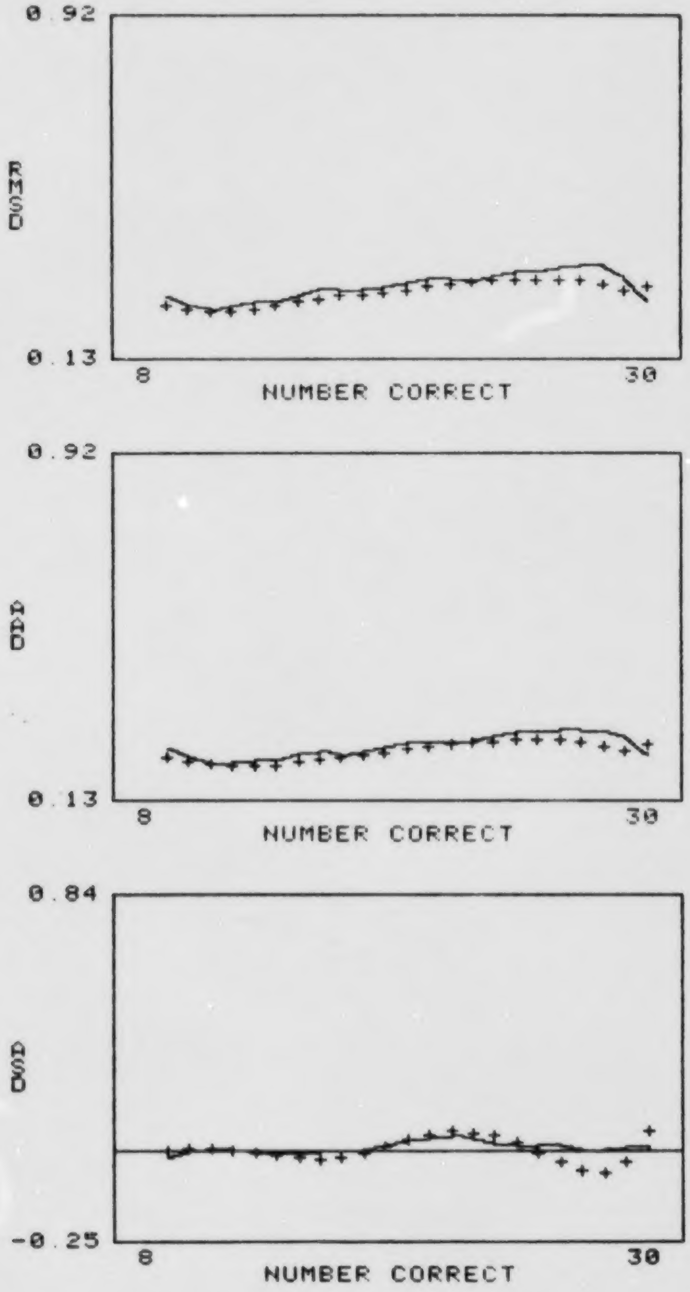


Figure B-63

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Postsmoothed by cubic smoothing splines

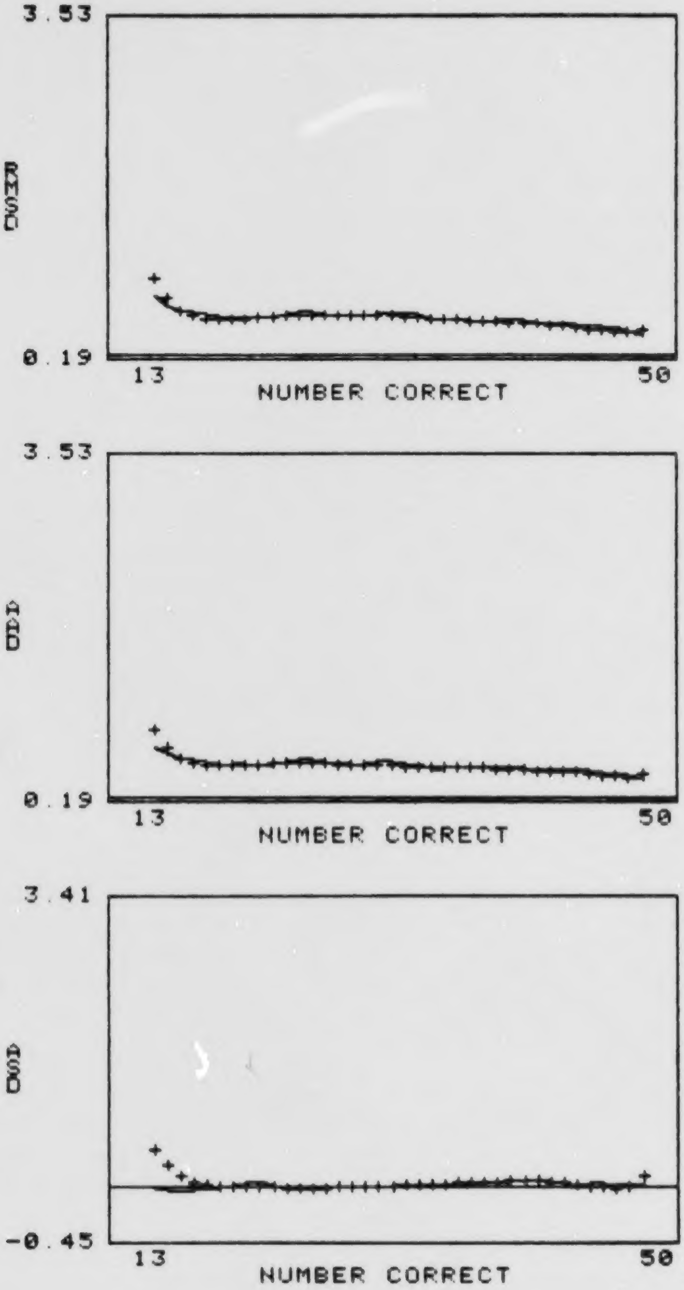


Figure B-64

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 20
Test Type: Operational
Smoothing: Postsmoothed by cubic smoothing splines

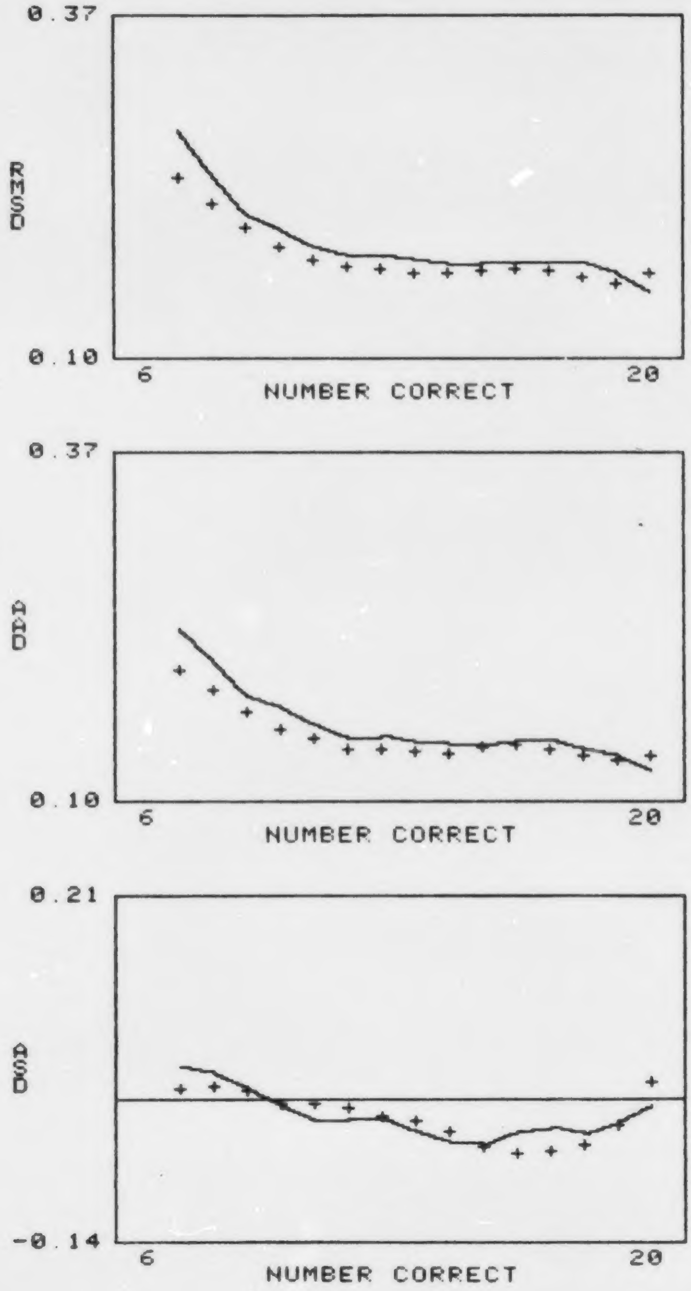


Figure B-65

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Postsmoothed by cubic smoothing splines

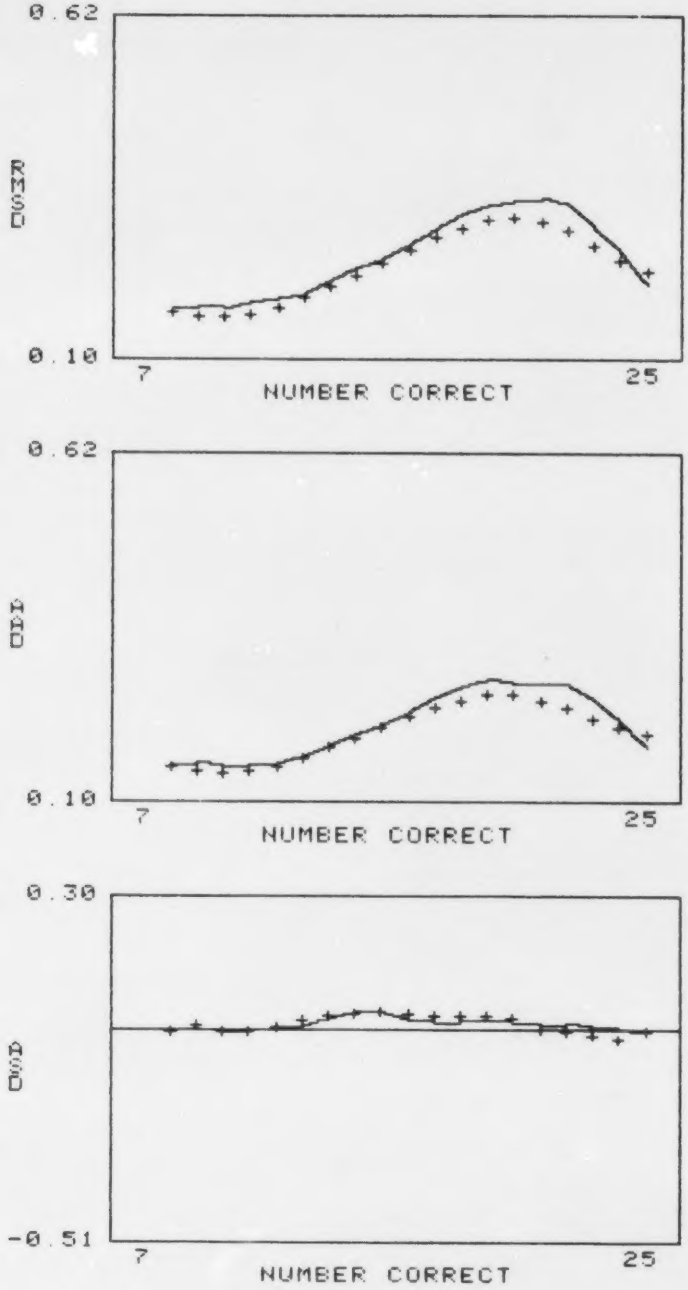


Figure B-66

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 15
Test Type: Simulated
Smoothing: Postsmoothed by 5-point moving weighted averages

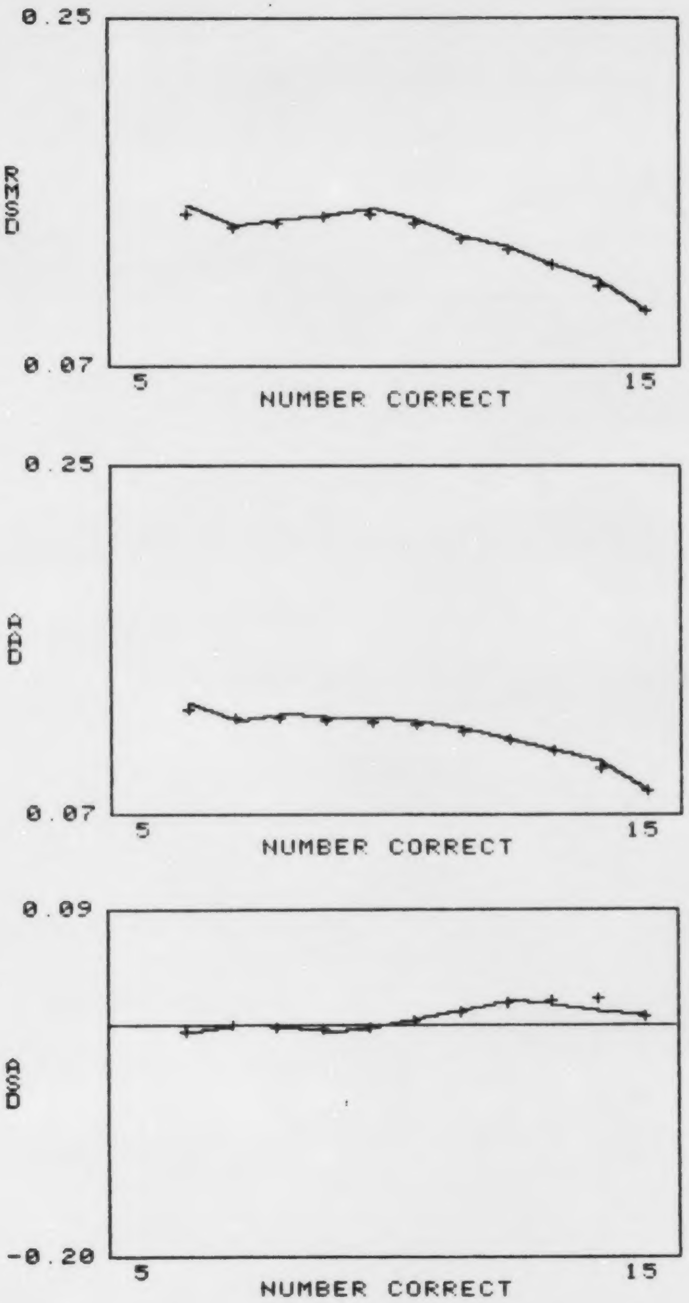


Figure B-67

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Postsmoothed by 5-point moving weighted averages

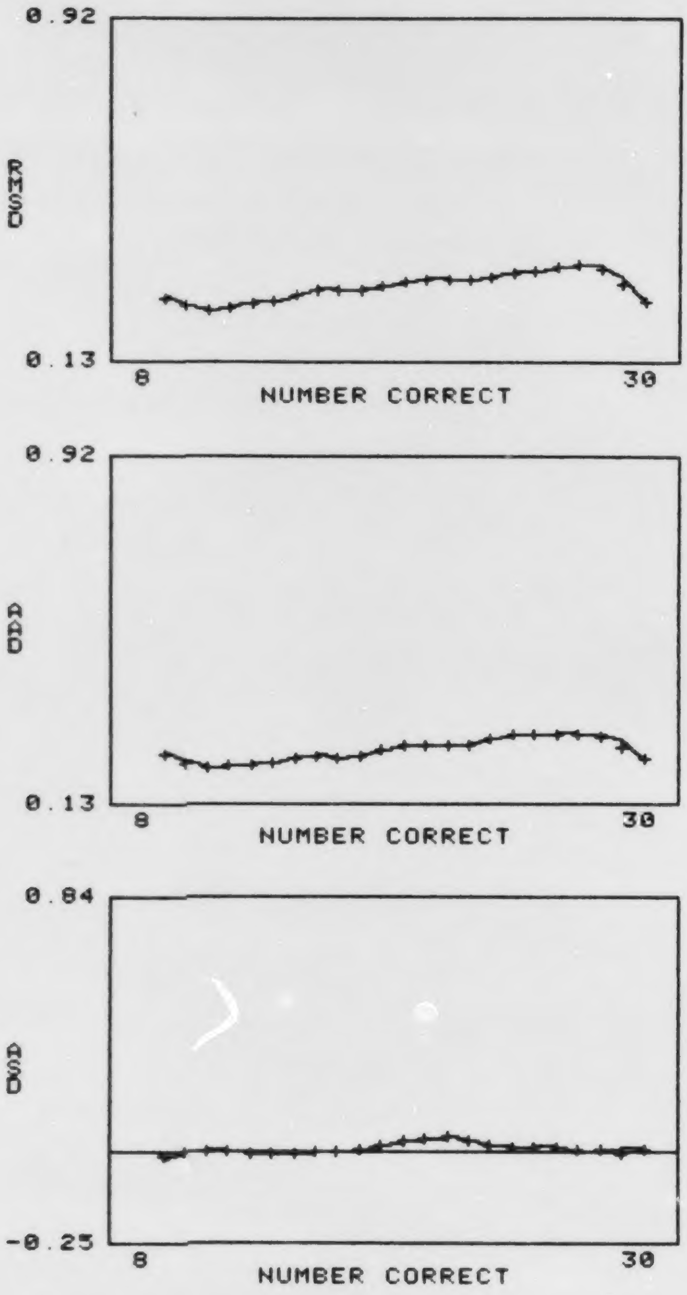


Figure B-68

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Postsmoothed by 5-point moving weighted averages

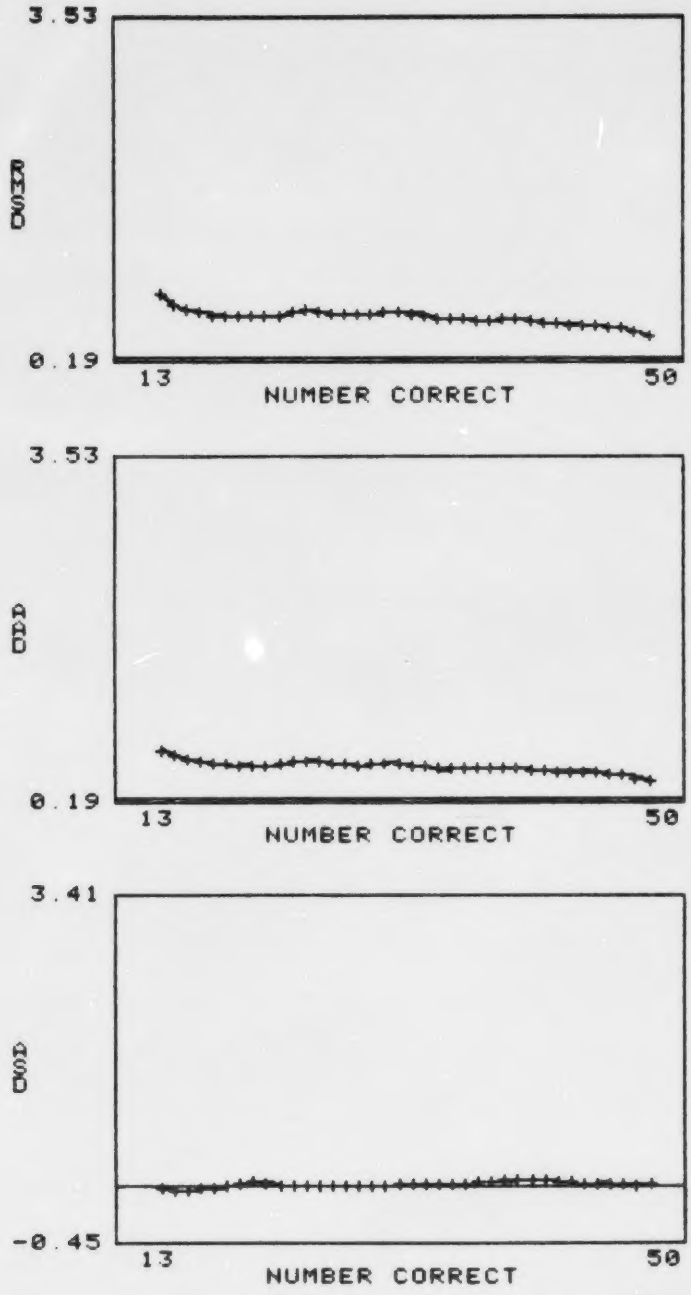


Figure B-69

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 20
Test Type: Operational
Smoothing: Postsmoothed by 5-point moving weighted averages

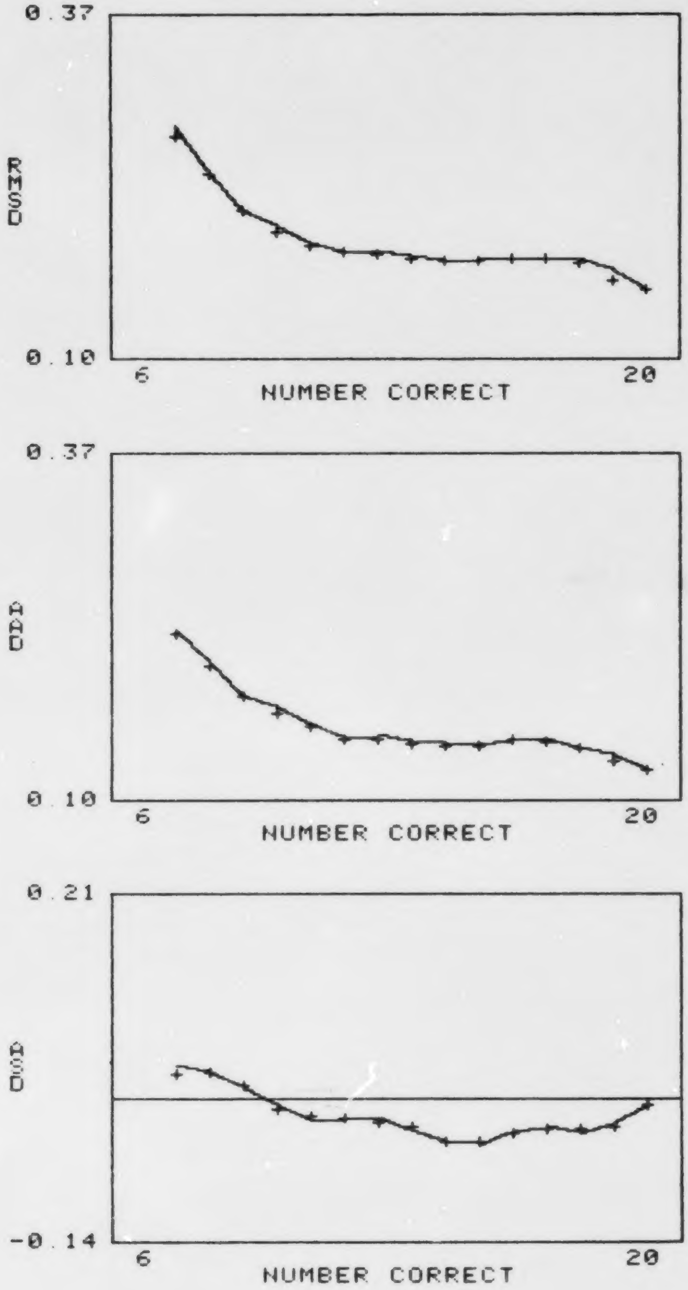


Figure B-70

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Postsmoothed by 5-point moving weighted averages

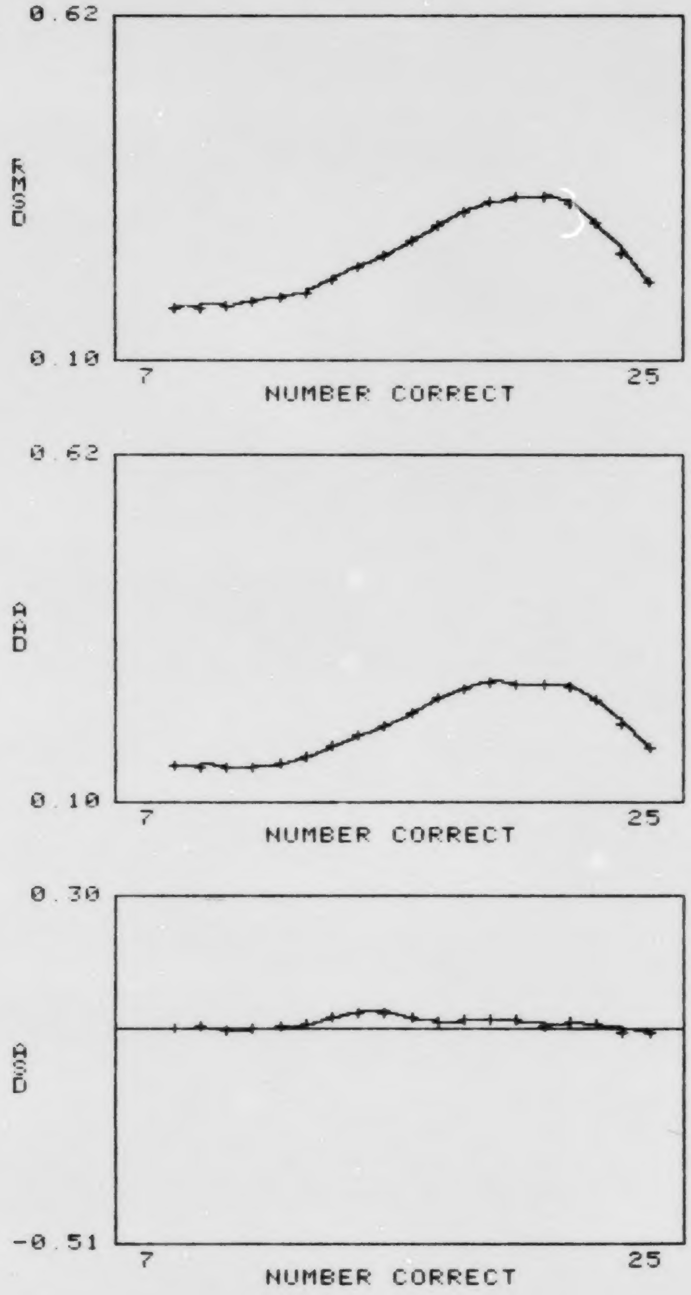


Figure B-71

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15

Test Type: Simulated

Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by orthogonal regression

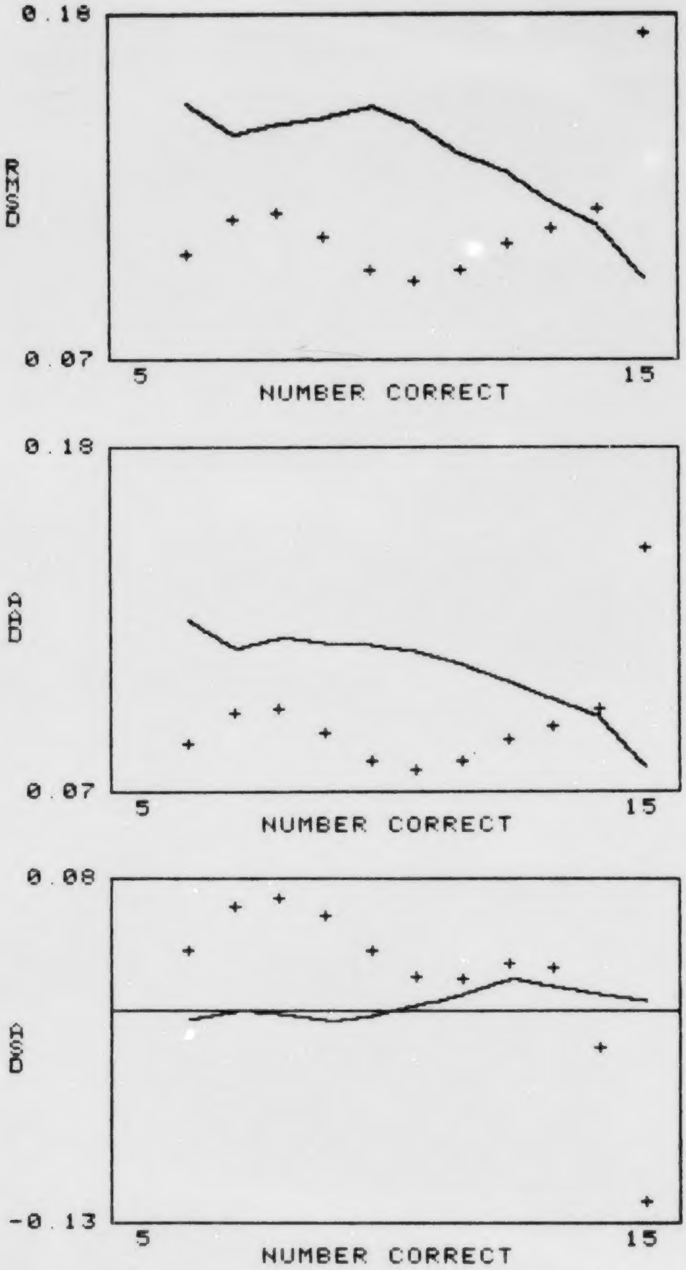


Figure B-72

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by orthogonal regression

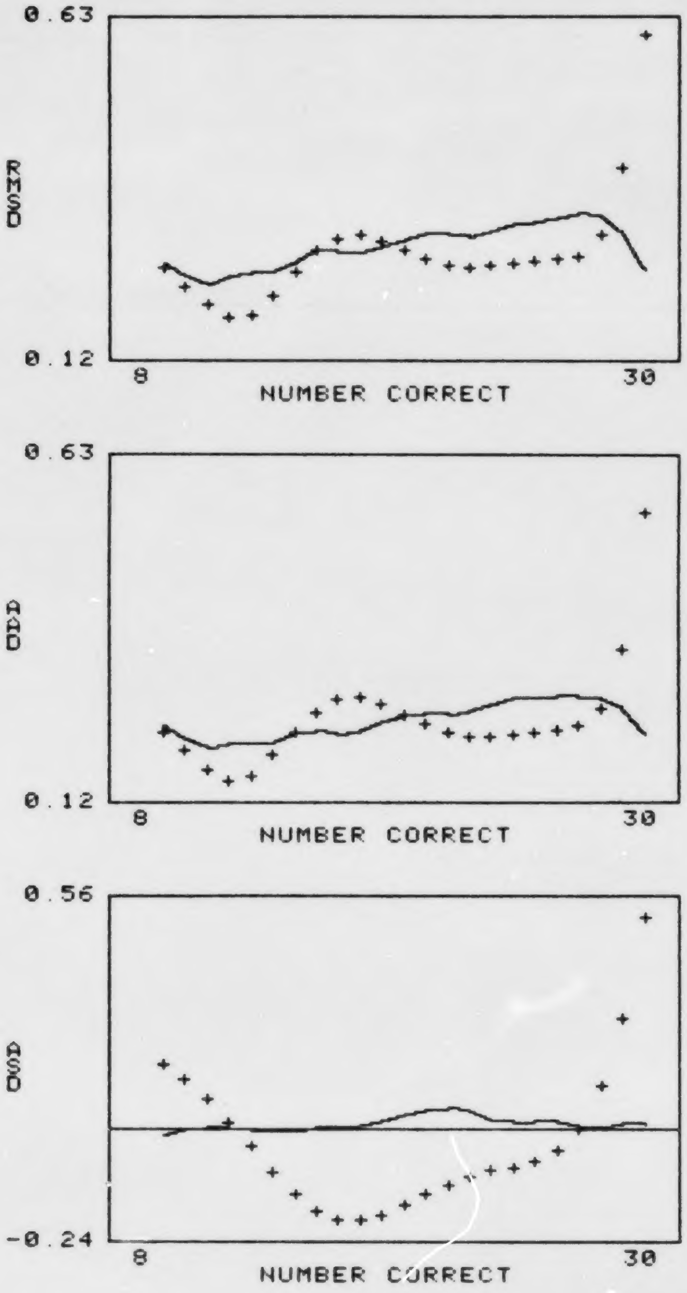


Figure B-73

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50

Test Type: Simulated

Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by orthogonal regression

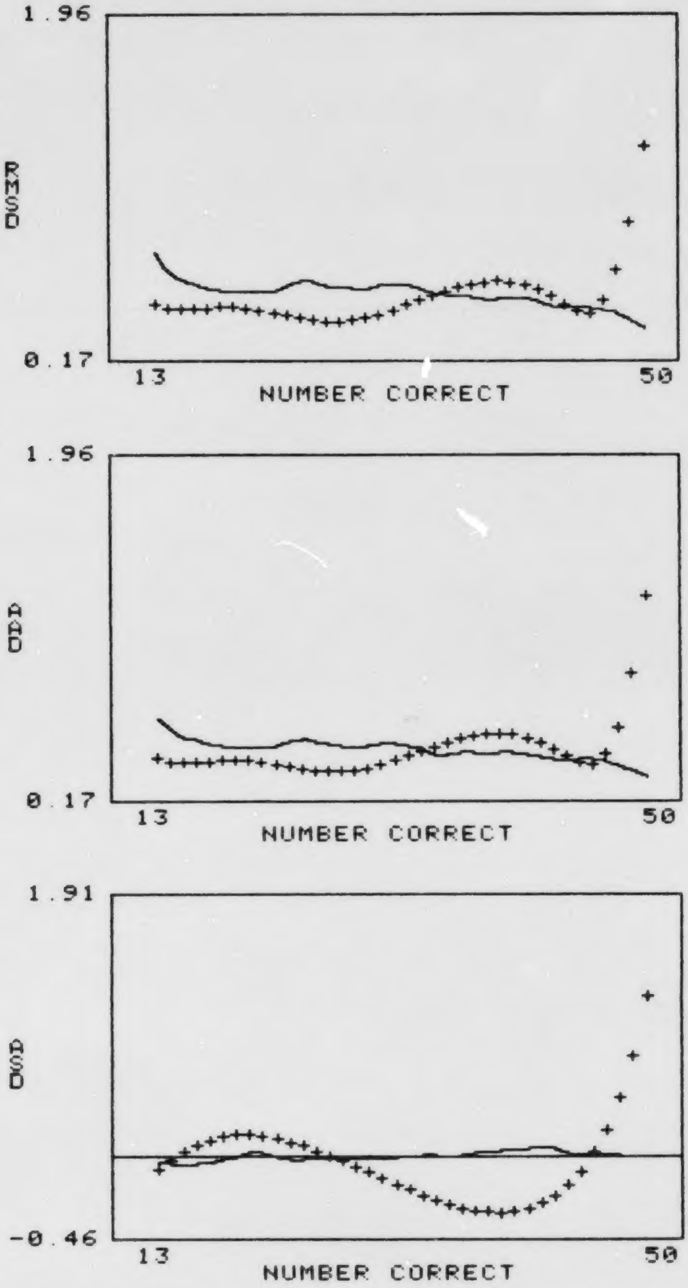


Figure B-74

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 20
Test Type: Operational
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by orthogonal regression

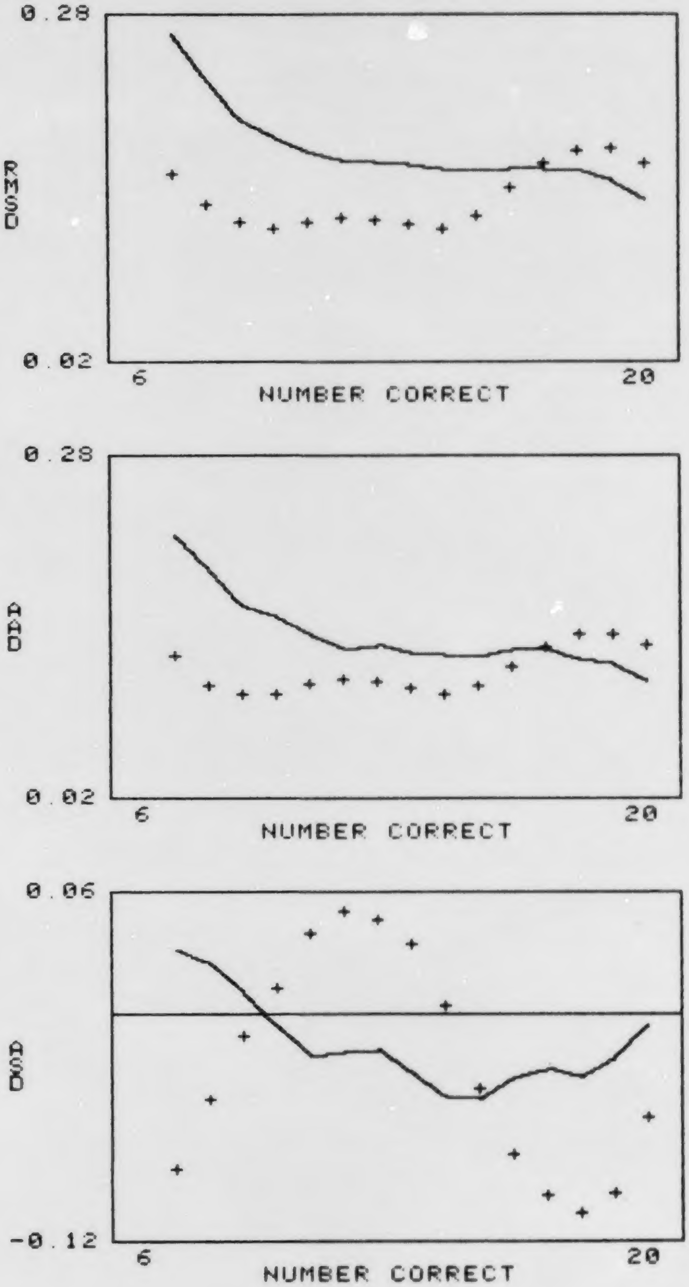


Figure B-75

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 25

Test Type: Operational

Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by orthogonal regression

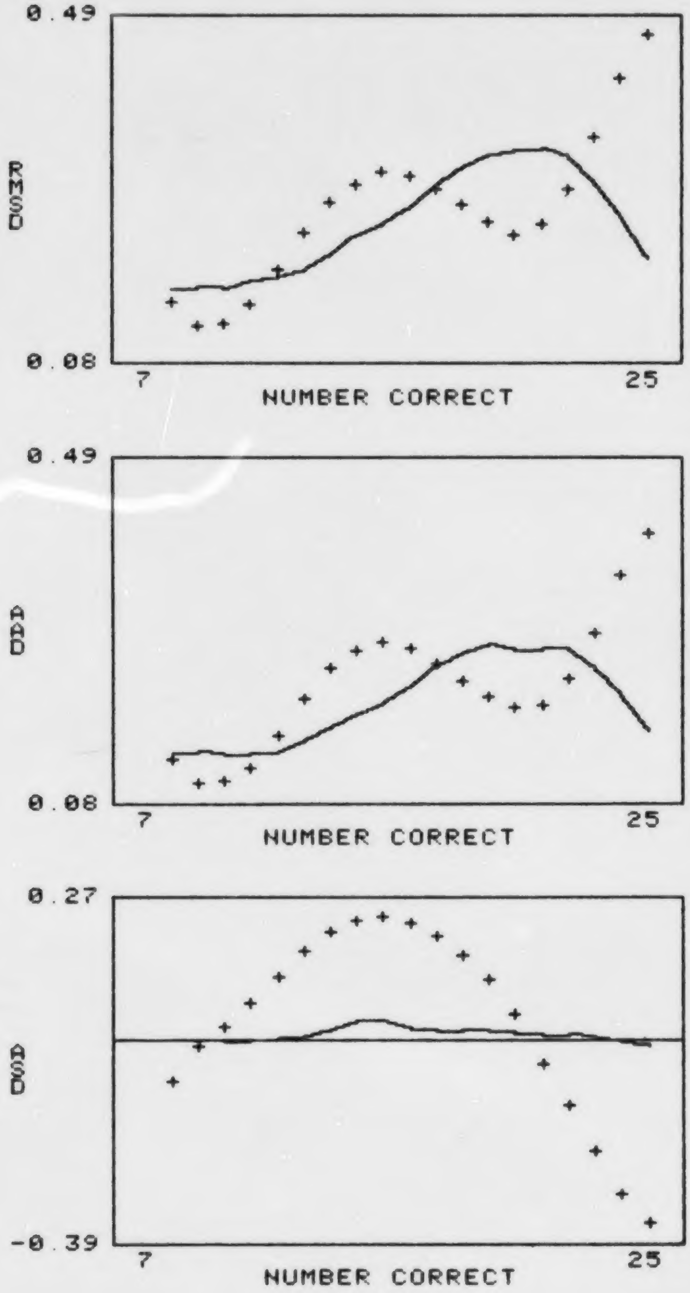


Figure B-76

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length:	15
Test Type:	Simulated
Smoothing:	Combined presmoothing by negative hypergeometric and postsmoothing by quadratic regression

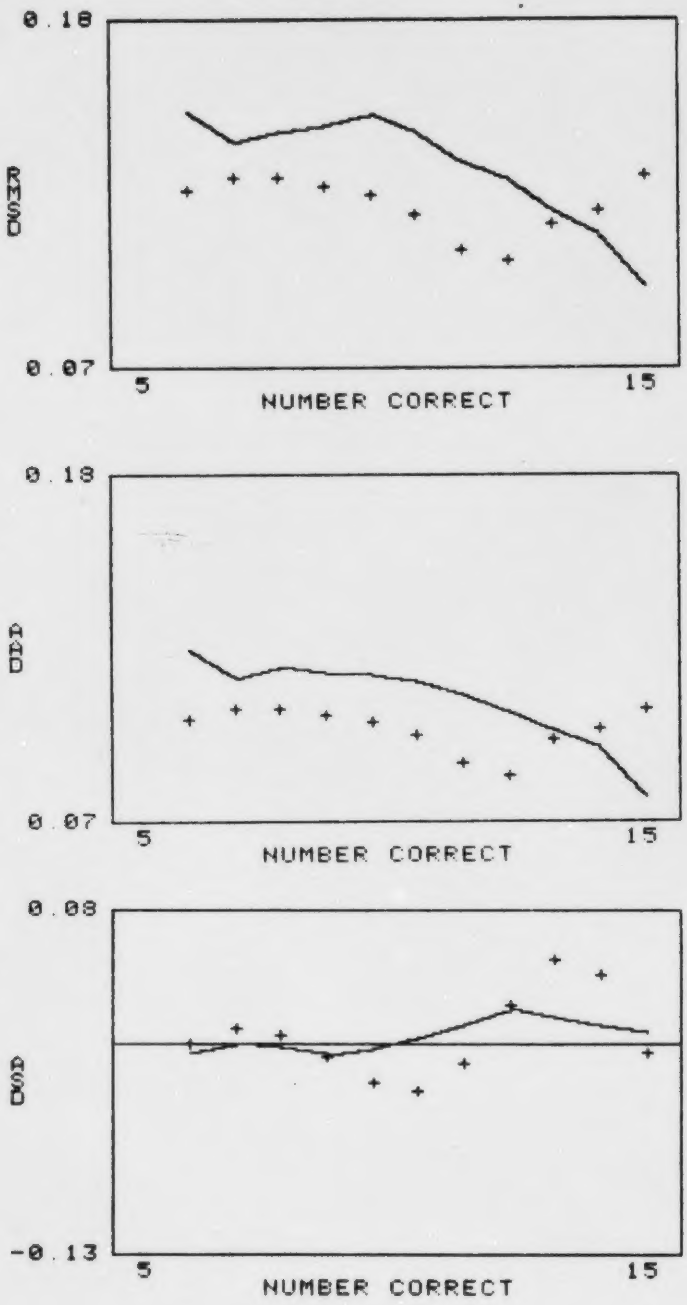


Figure B-77

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 30
Test Type: Simulated
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by quadratic regression

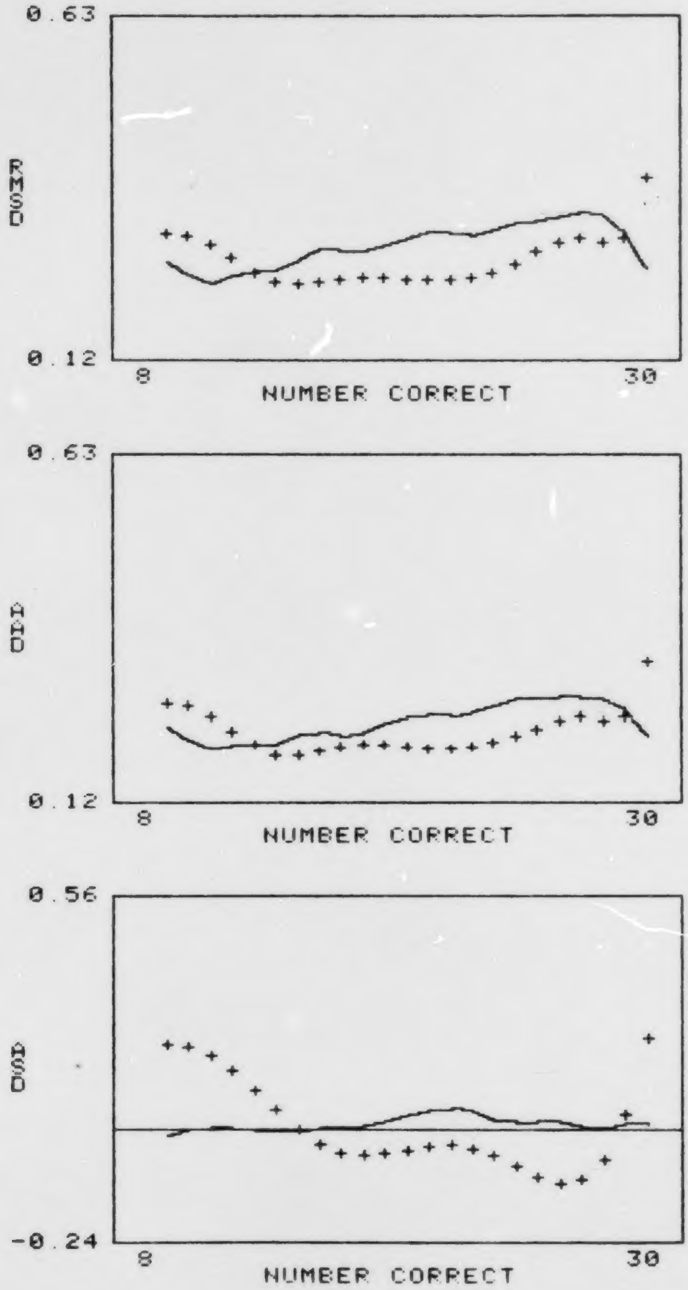


Figure B-78

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by quadratic regression

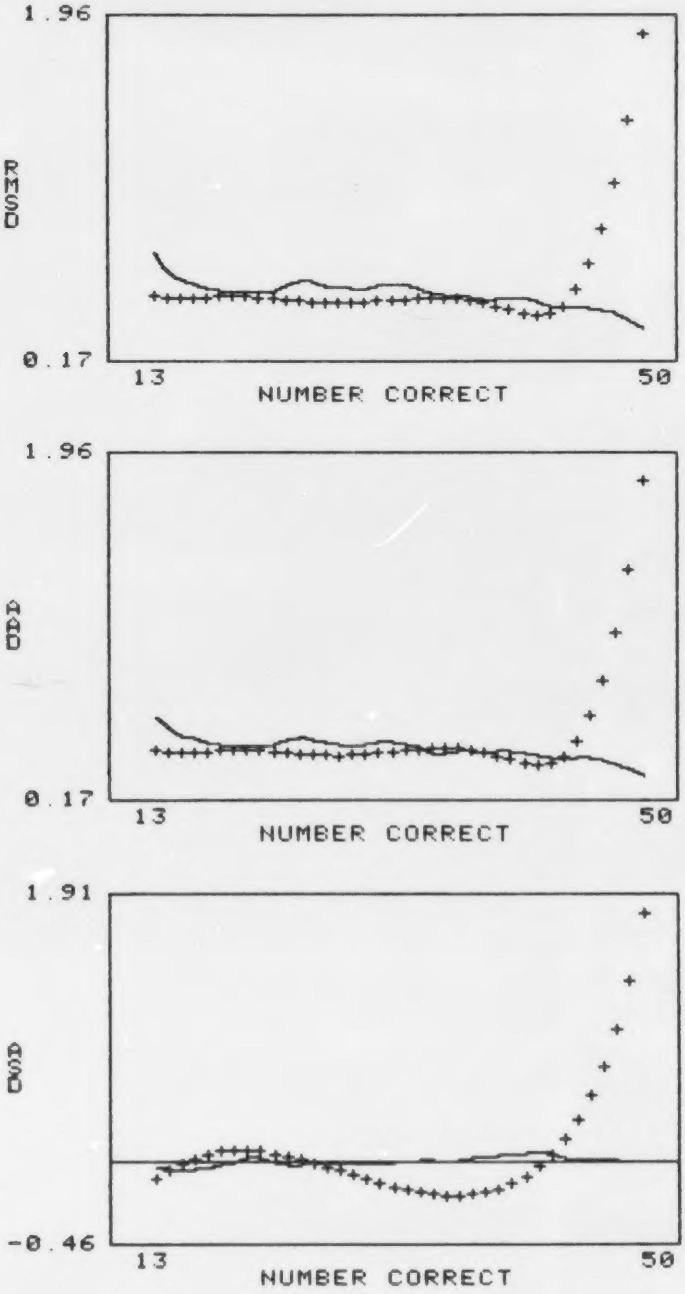


Figure B-79

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 20

Test Type: Operational

Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by quadratic regression

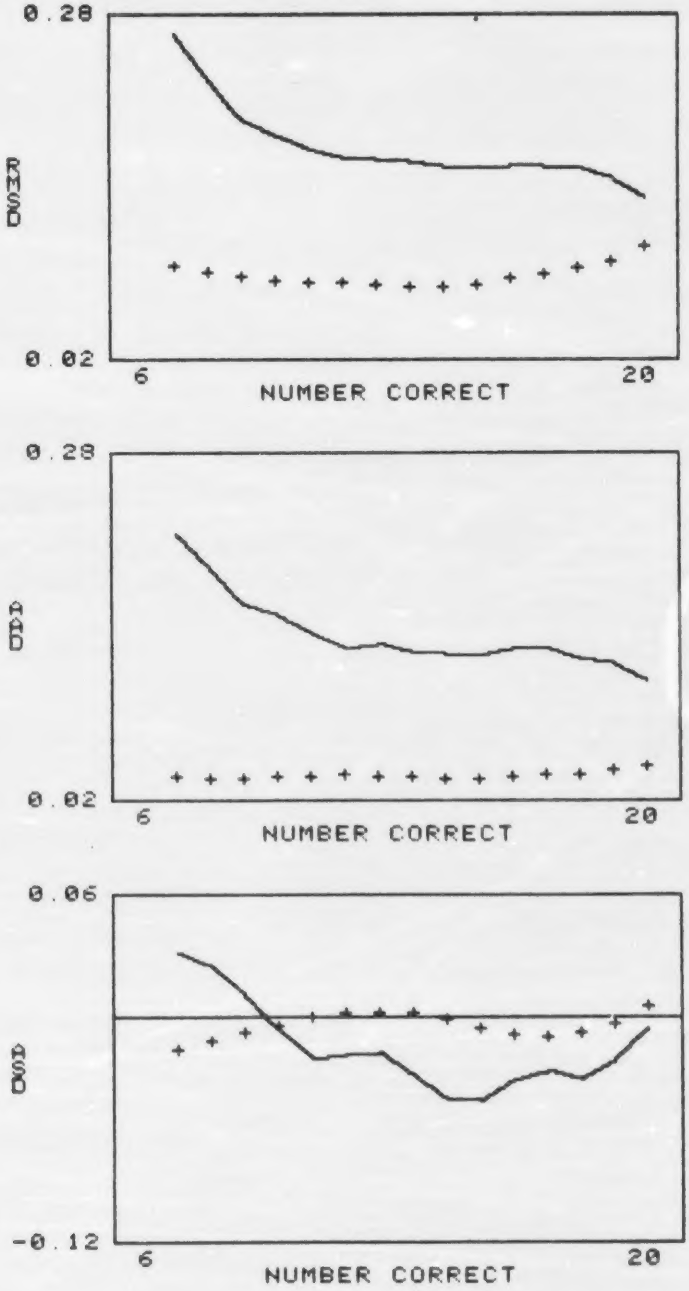


Figure B-80

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.
Test Length: 25
Test Type: Operational
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by quadratic regression

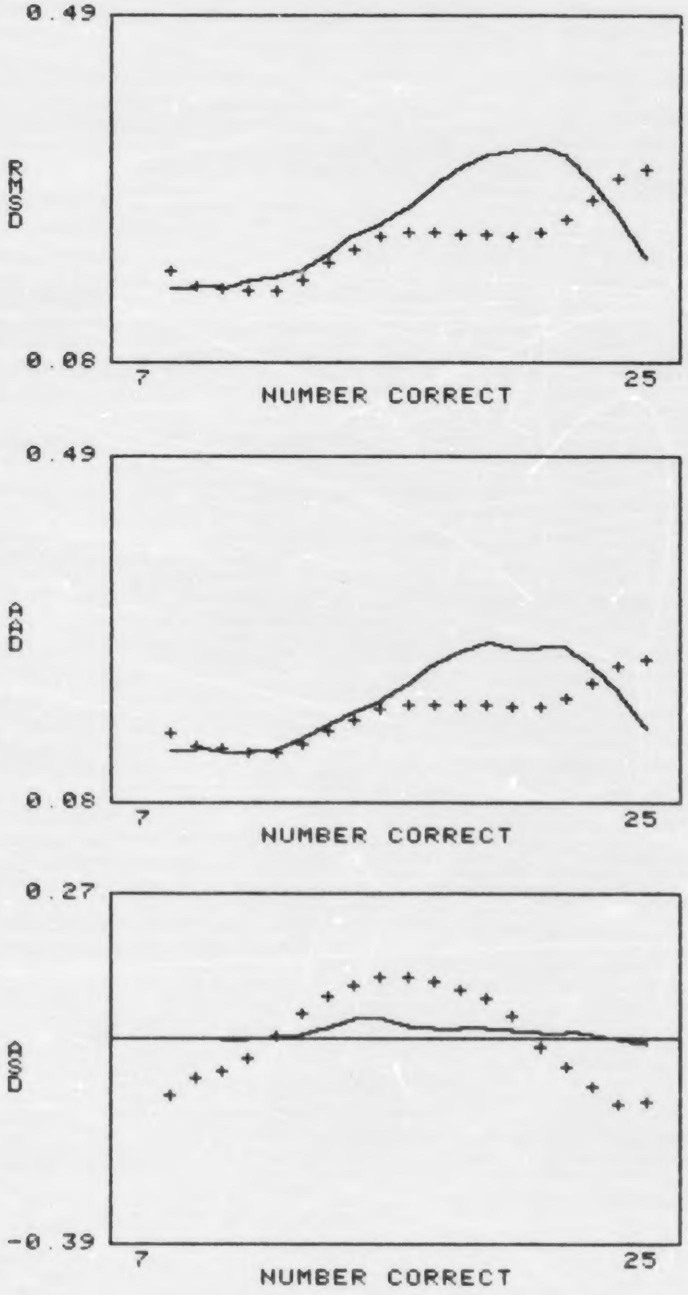


Figure B-81

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15

Test Type: Simulated

Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by 5-point moving weighted averages

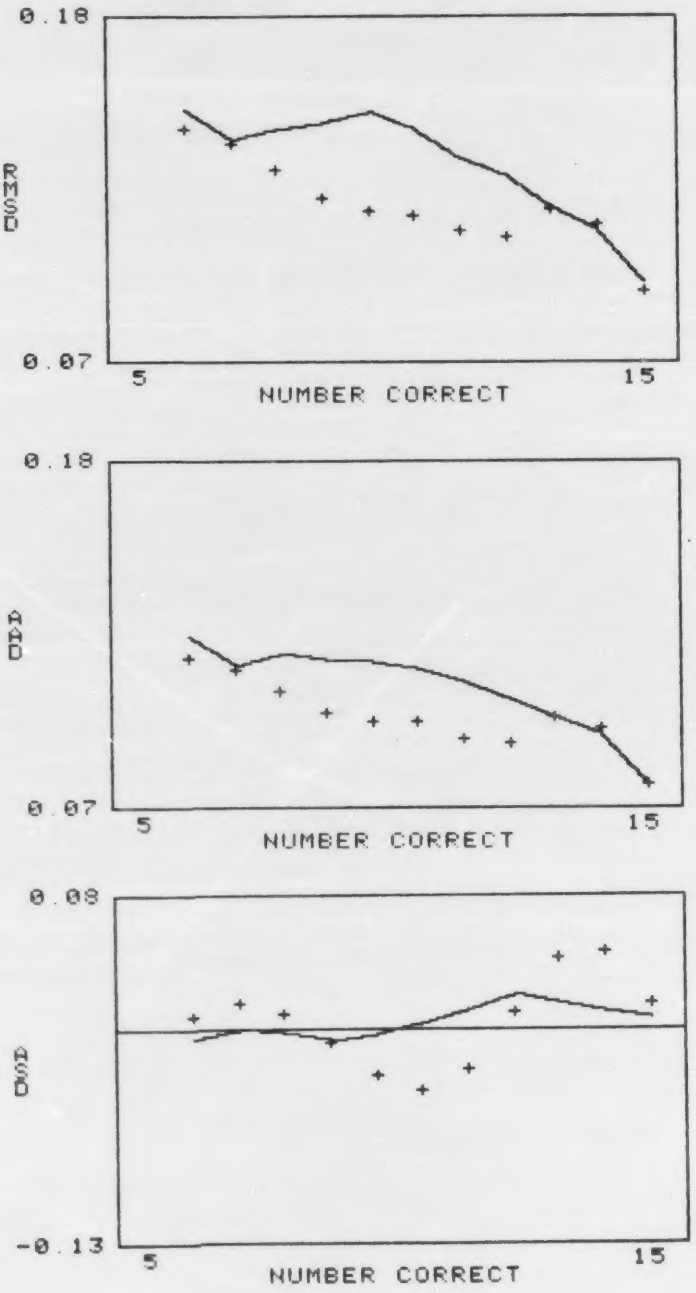


Figure B-82

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length:	30
Test Type:	Simulated
Smoothing:	Combined presmoothing by negative hypergeometric and postsmoothing by 5-point moving weighted averages

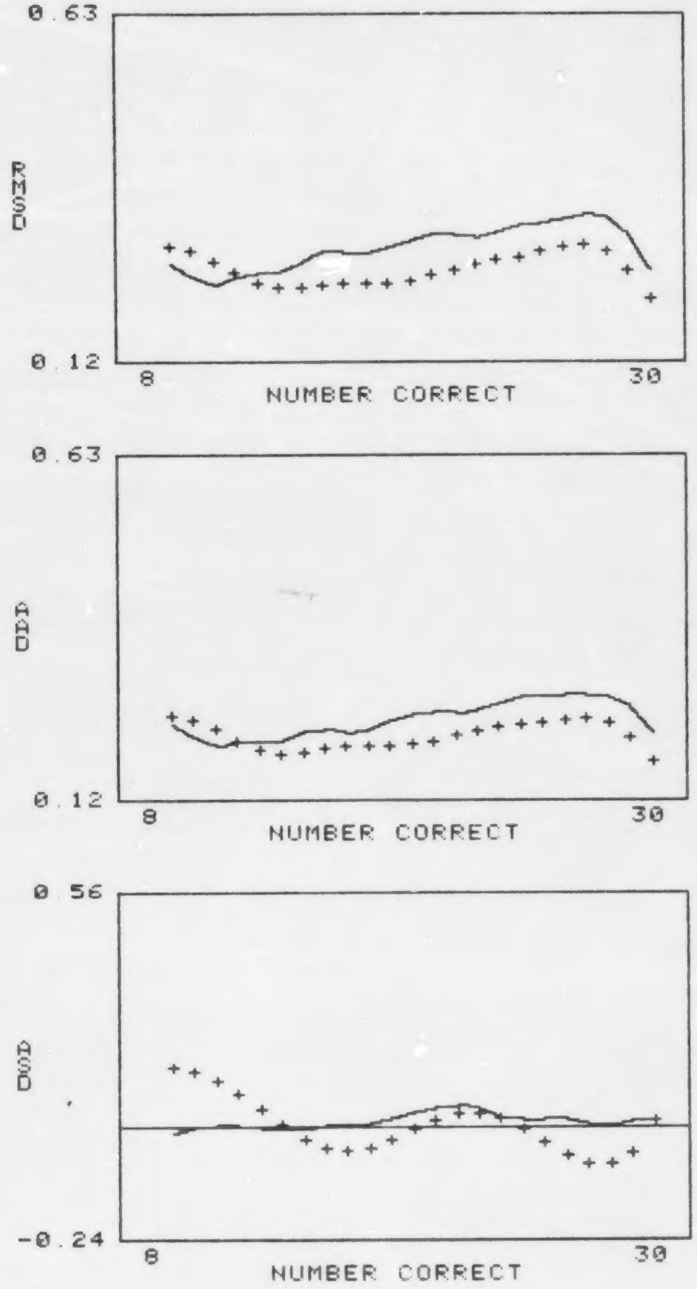


Figure B-83

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50

Test Type: Simulated

Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by 5-point moving weighted averages

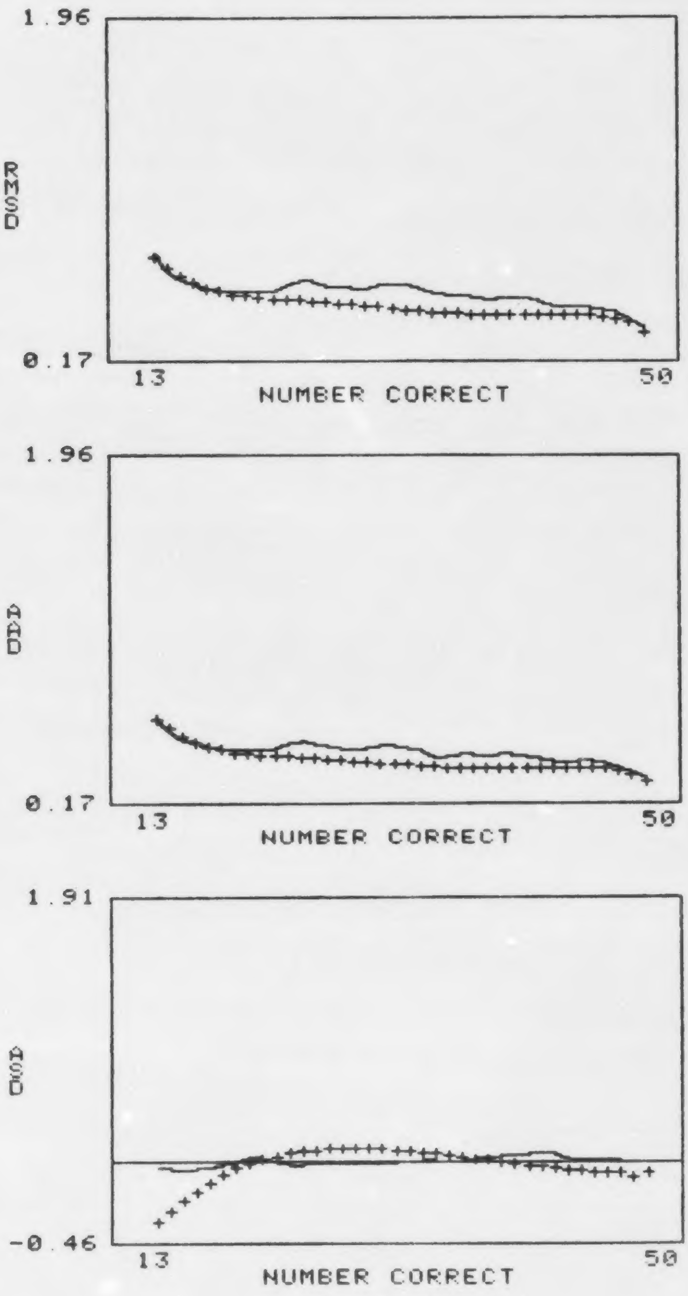


Figure B-64

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length:	20
Test Type:	Operational
Smoothing:	Combined presmoothing by negative hypergeometric and postsmoothing by 5-point moving weighted averages

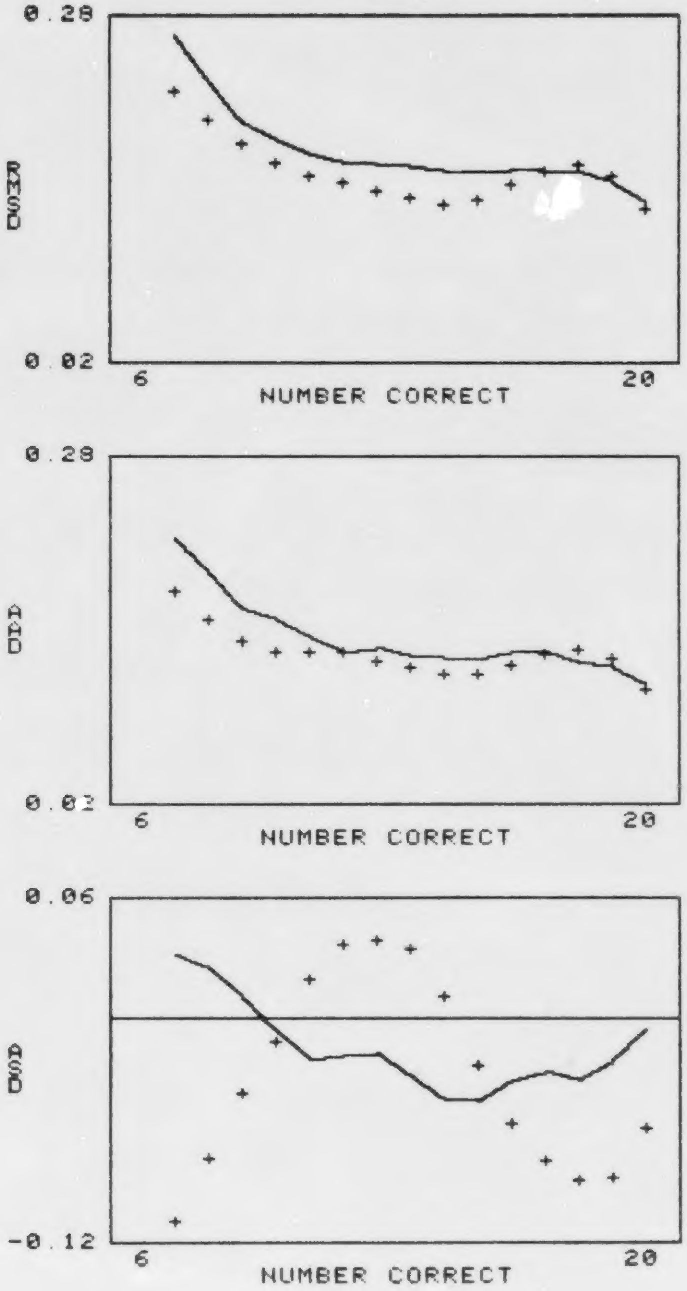


Figure B-85

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length:	**25**
Test Type:	**Operational**
Smoothing:	**Combined presmoothing by negative hypergeometric and postsmoothing by 5-point moving weighted averages**

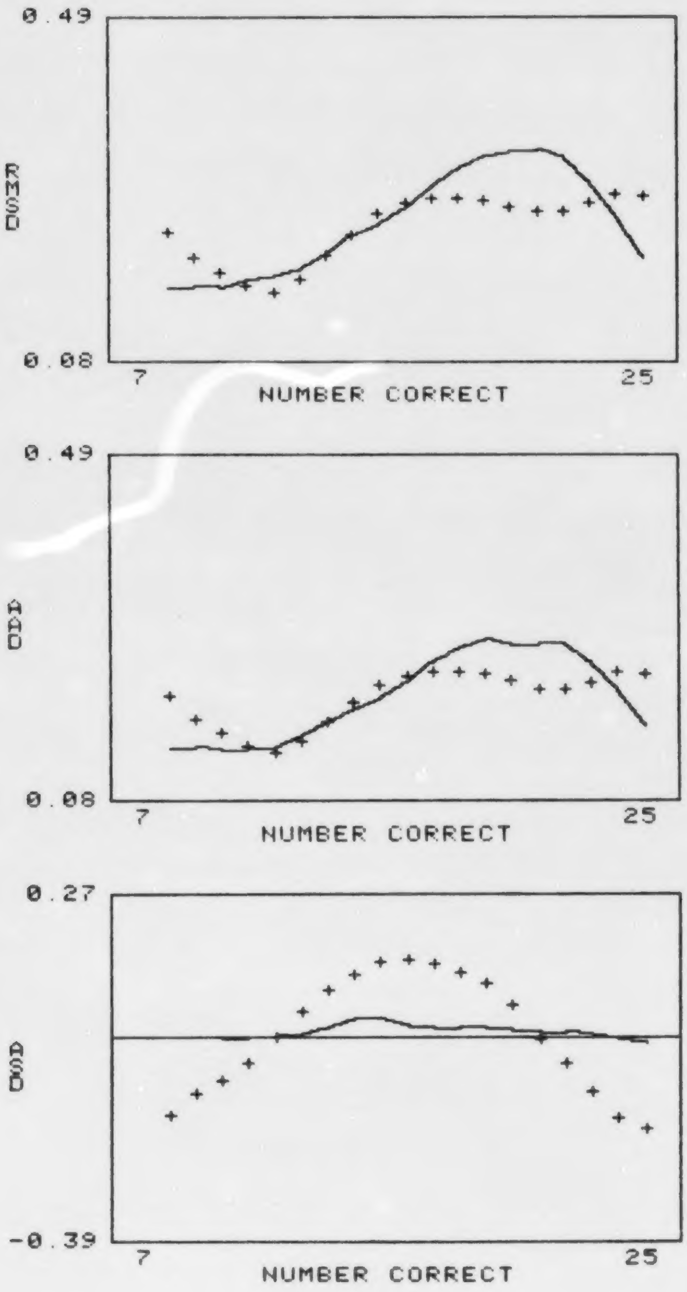


Figure B-86

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15

Test Type: Simulated

Smoothing: Combined presmoothing by 3-point moving weighted averages and postsmoothing by 5-point moving weighted averages

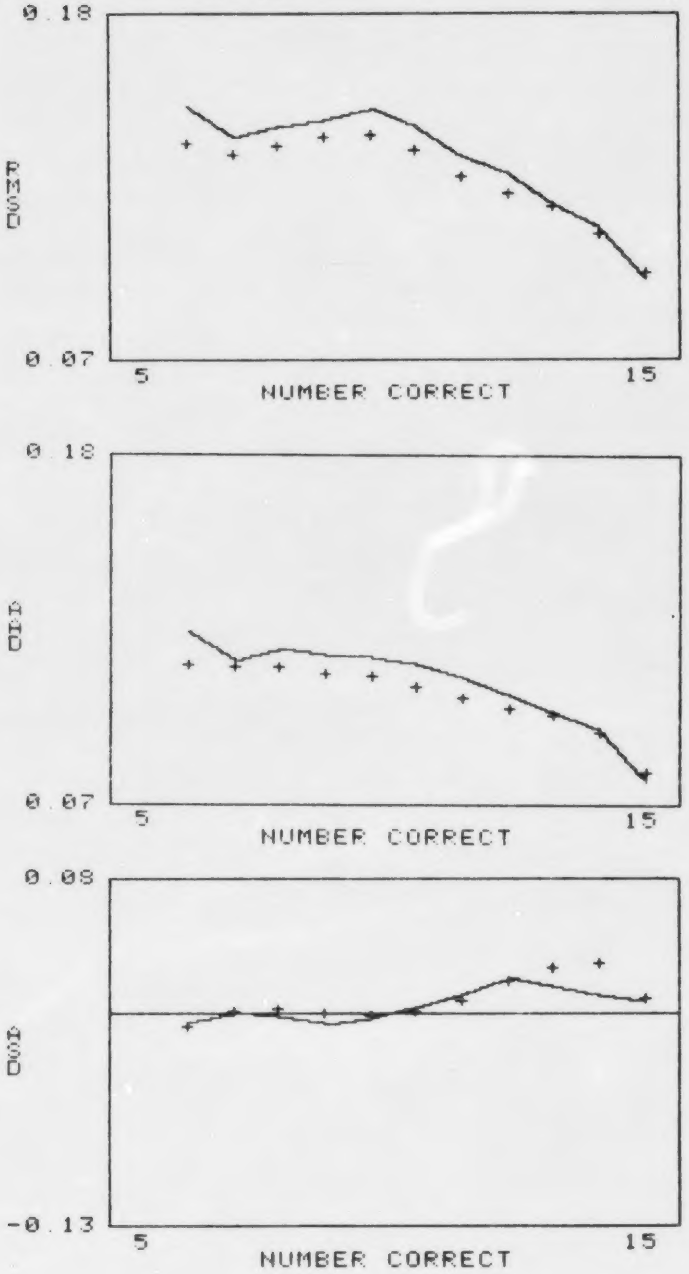


Figure B-87

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Combined presmoothing by 3-point moving weighted averages and postsmoothing by 5-point moving weighted averages

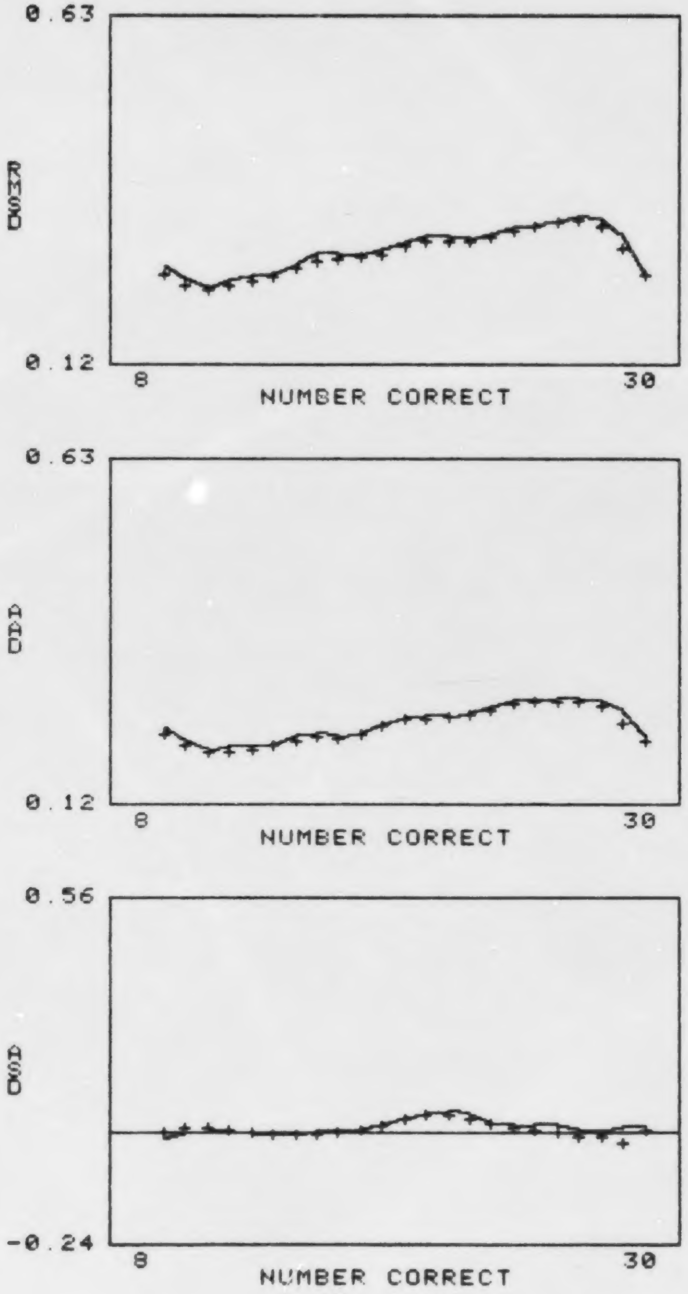


Figure B-88

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50

Test Type: Simulated

Smoothing: Combined presmoothing by 3-point moving weighted averages and postsmoothing by 5-point moving weighted averages

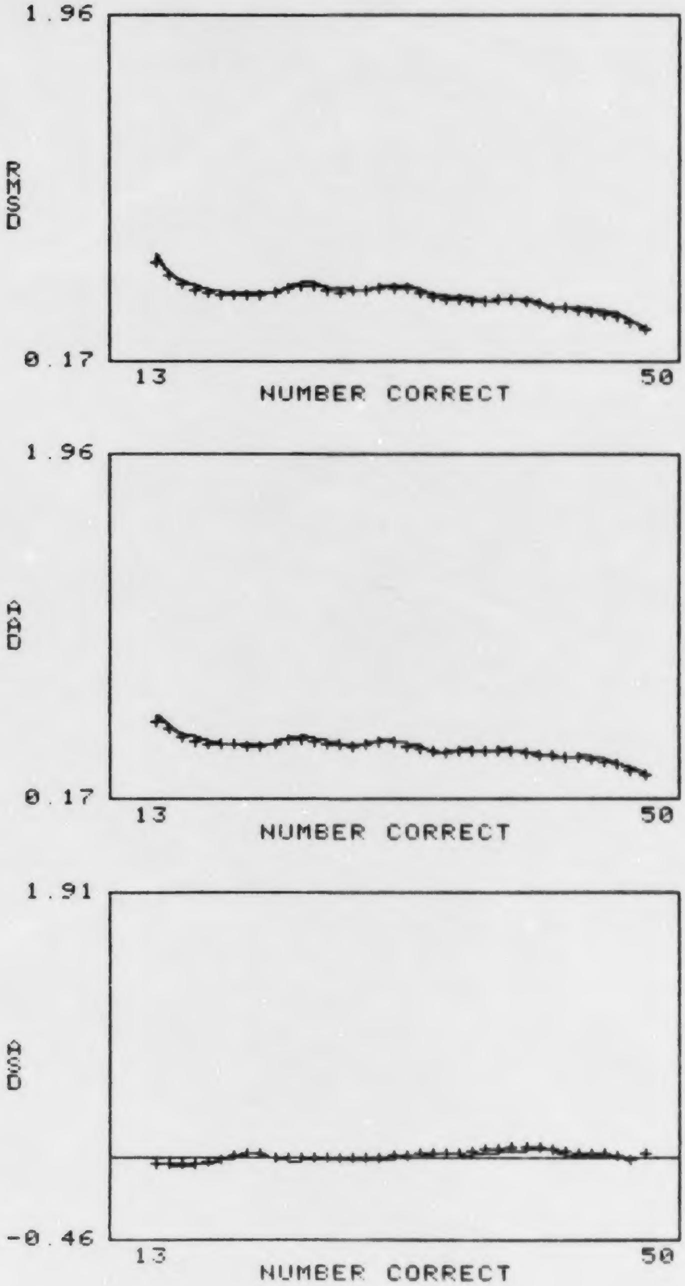


Figure B-89

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + mark).

Test Length:	20
Test Type:	Operational
Smoothing:	Combined presmoothing by 3-point moving weighted averages and postsmoothing by 5-point moving weighted averages

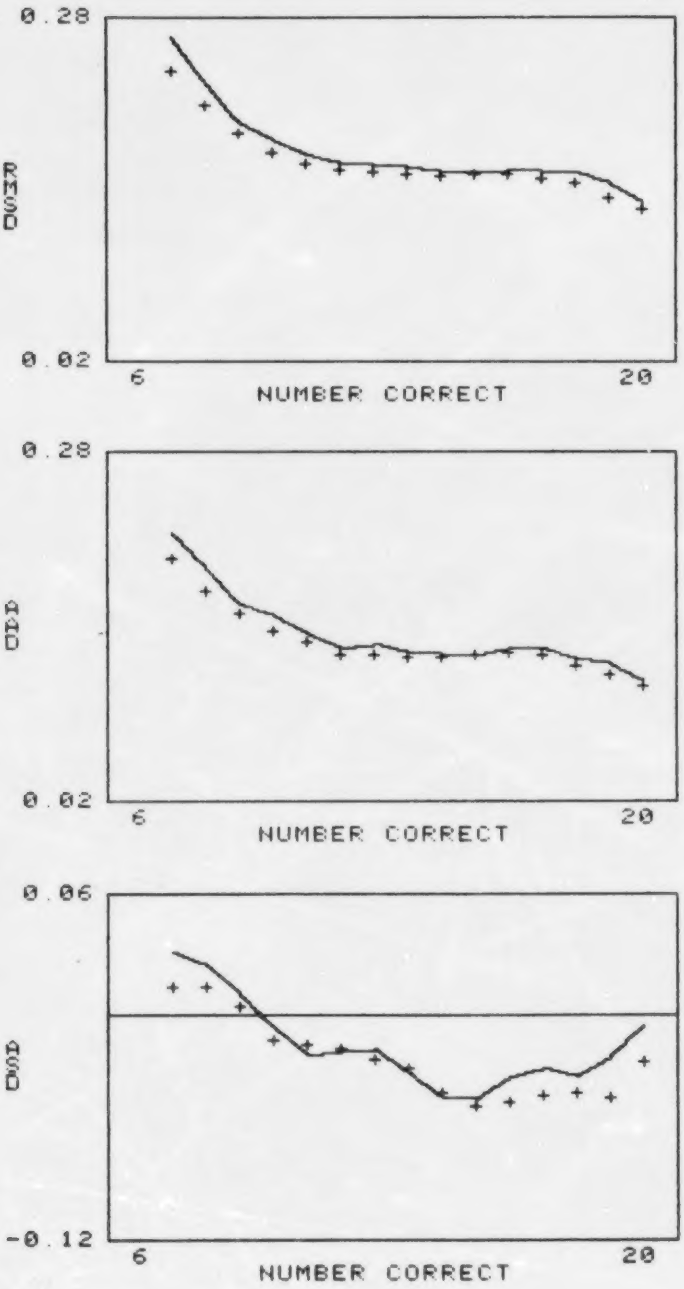


Figure B-90

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + mark).

Test Length: 25
Test Type: Operational
Smoothing: Combined presmoothing by 3-point moving weighted averages and postsmoothing by 5-point moving weighted averages

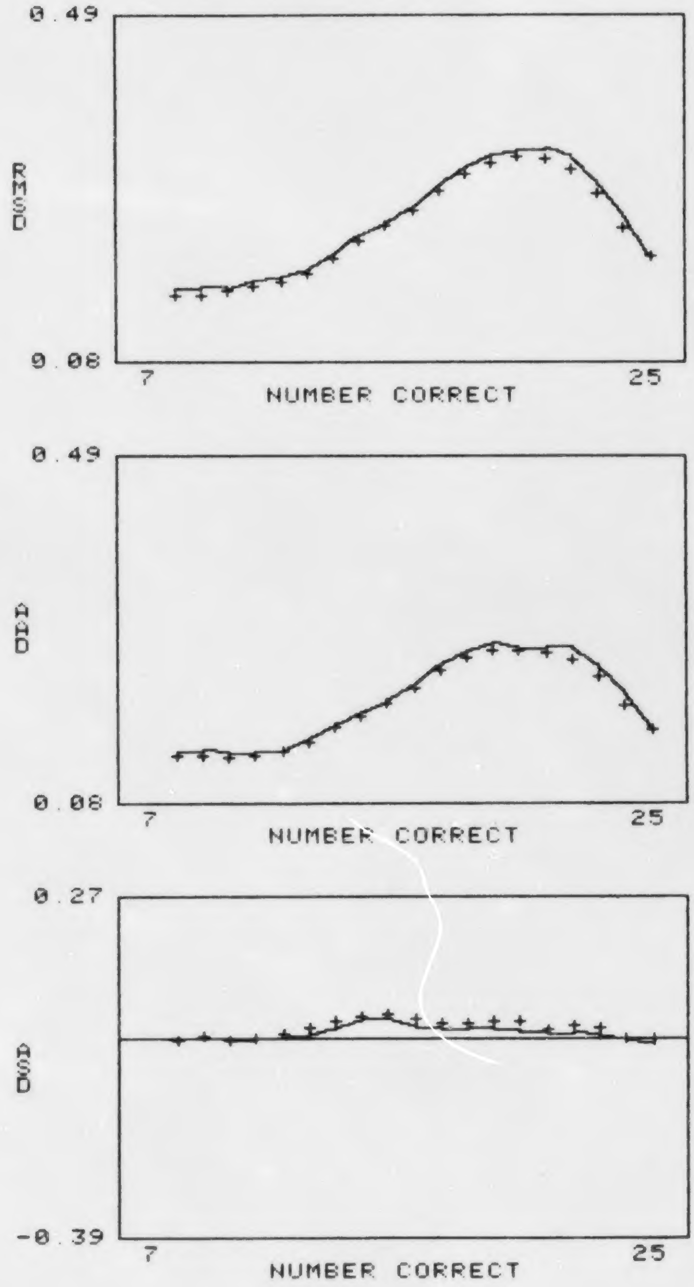


Figure B-91

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 15
Test Type: Simulated
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by cubic splines

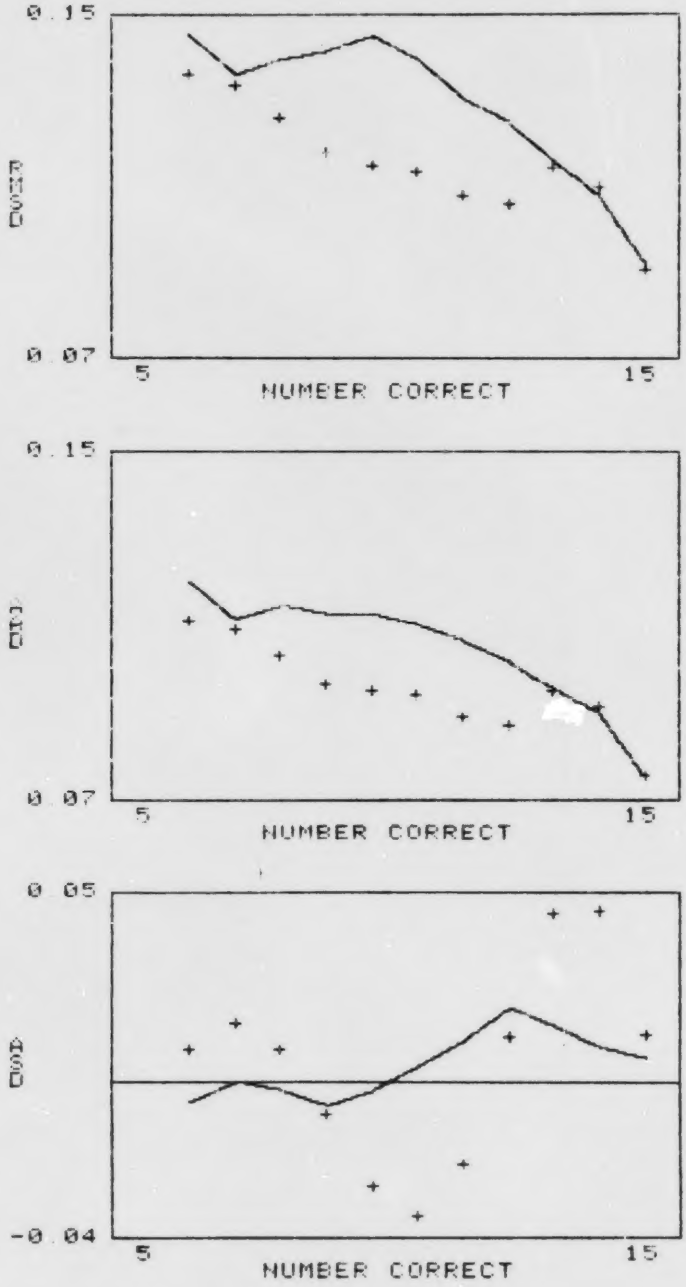


Figure B-92

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 30
Test Type: Simulated
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by cubic splines

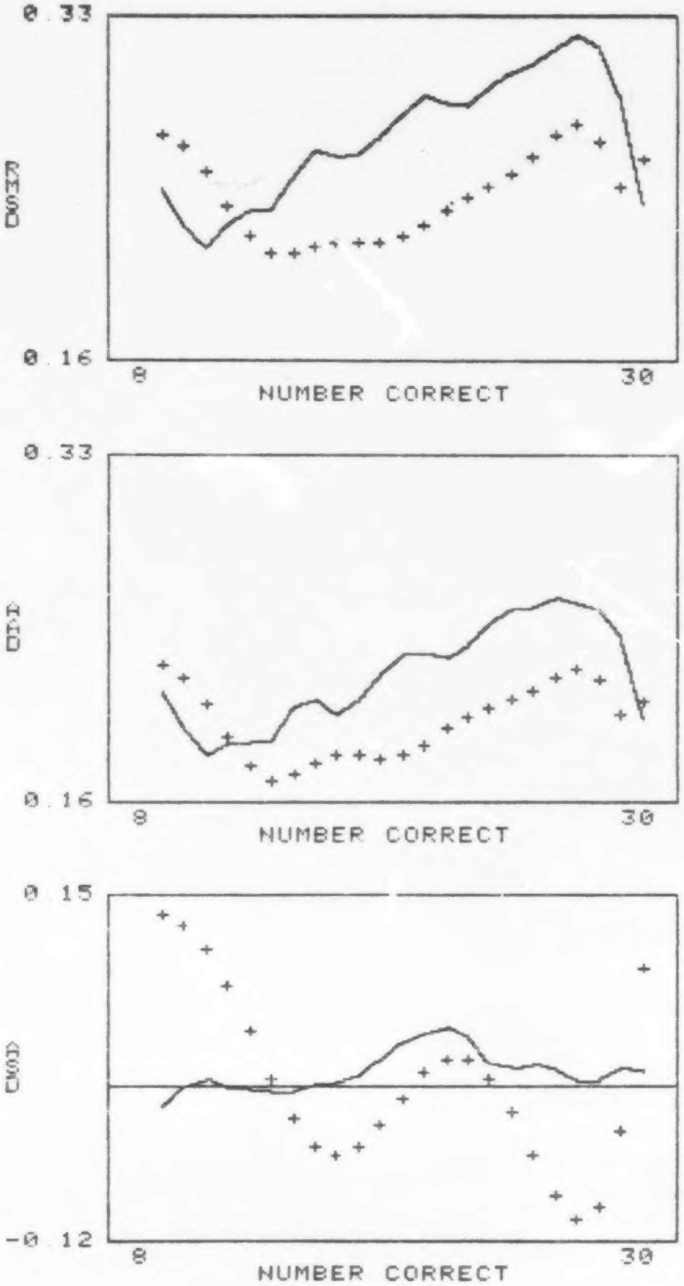


Figure B-93

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + marks.

Test Length: 50
Test Type: Simulated
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by cubic splines

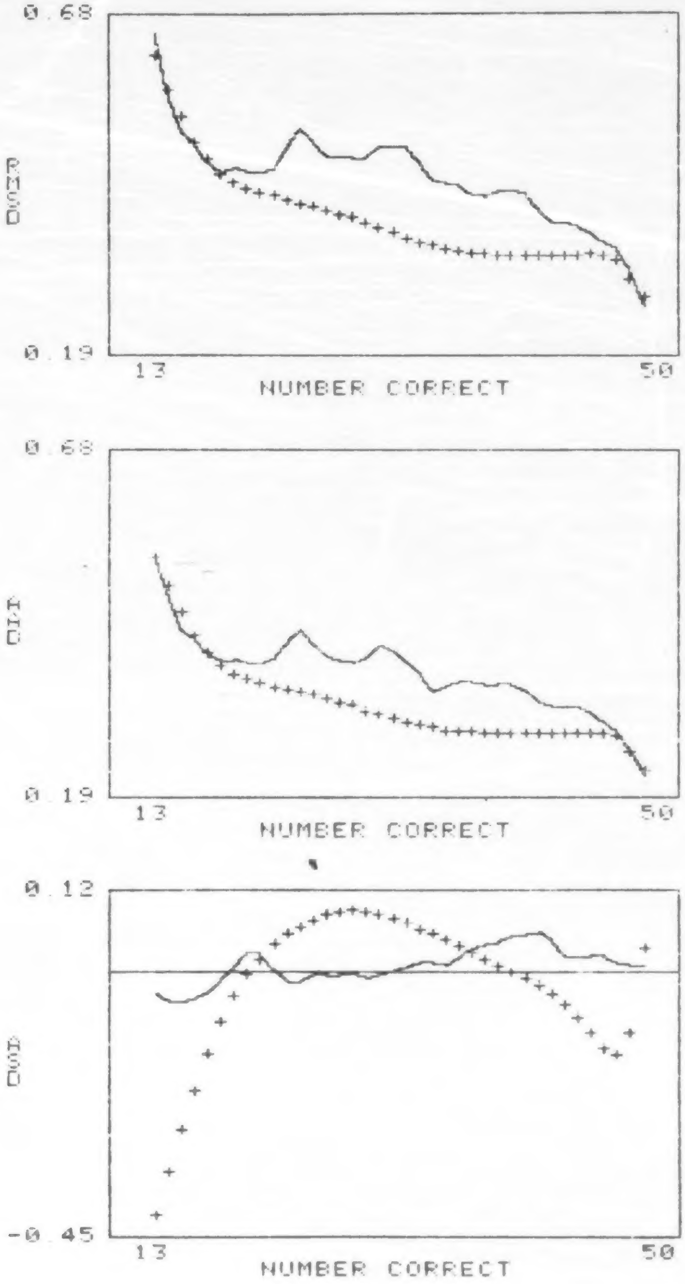


Figure B-94

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + mark).

Test Length: 20

Test Type: Operational

Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by cubic splines

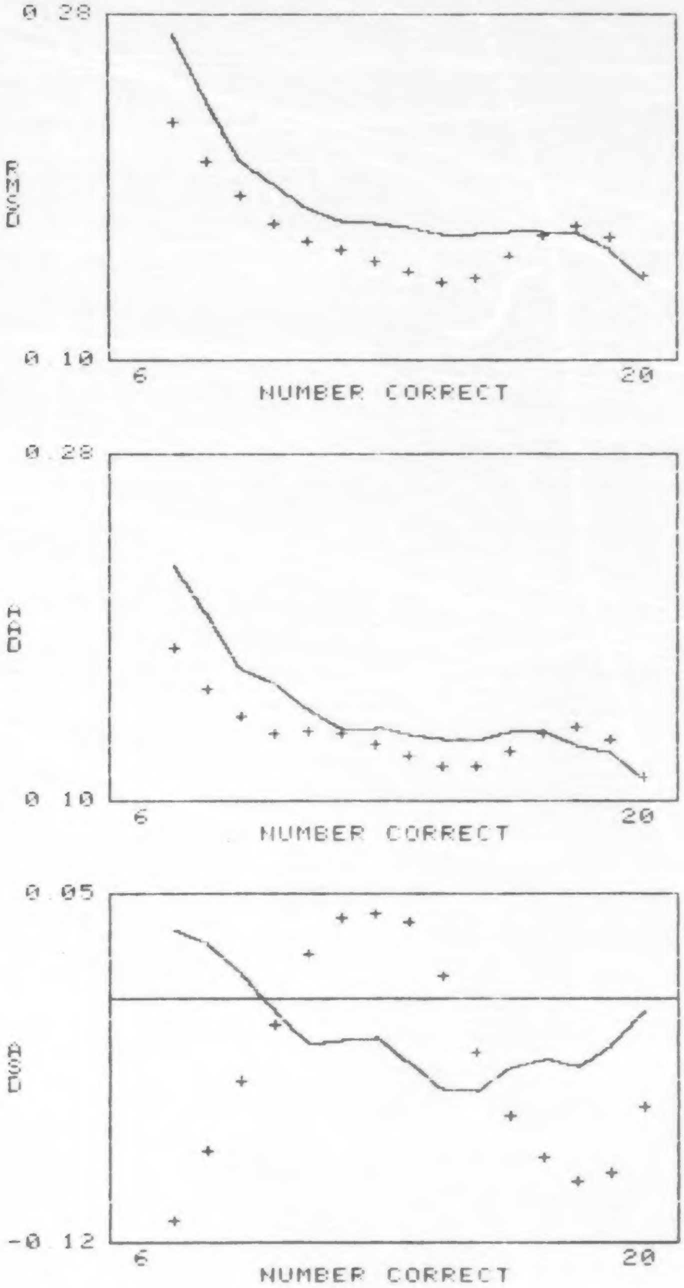


Figure B-95

Deviations of sample equatings (RMSD, AAD, and ASD) from criterion equating. Unsmoothed equating: solid line; smoothed equating: + mark).

Test Length: 25
Test Type: Operational
Smoothing: Combined presmoothing by negative hypergeometric and postsmoothing by cubic splines